Journal of Political Economy

Volume 85 Number 4 August 1977

iii In Memoriam: Harry G. Johnson

Articles

661 Financial Flow Variables and the Short-Run Determination of Long-Term Interest Rates
Benjamin M. Friedman

691 The Demand for Quality-adjusted Cash Balances: Price Uncertainty in the U.S. Demand for Money Function
Benjamin Klein

717 The Competitive Effects of Trust-Busting: A Portfolio Analysis
Malcolm R. Burns

741 Capital Punishment and Deterrence: Some Further Thoughts and Additional Evidence
Isaac Ehrlich

789 The Demand for Supplementary Health Insurance, or Do Deductibles Matter?
Emmett B. Keeler, Daniel T. Morrow, and Joseph P. Newhouse

803 Probability and Utility Estimates for Racetrack Bettors
Mukhtar M. Ali

817 The Variability of Expectations in Hyperinflations
Mohsin S. Khan

829 On the Structure and Stability of Political Markets
W. Mark Crain

843 On Uncertain Lifetimes
Robert J. Barro and James W. Friedman

851 Consumer Horizon: Further Evidence
Khan A. Mohabbat and Evangelos O. Simos

859 Equilibrium in Stable Markets
James A. Ohlson

865 On the Shadow Pricing of Traded Commodities
Peter G. Warr

SV02
Accession Number 81606
Date 26.5.81
A.M

Miscellany

873 An Example of a Trading Economy with Three Competitive Equilibria
L. S. Shapley and M. Shubik

876 Editors' Notice

Book Reviews

877 Jagdish N. Bhagwati and T. N. Srinivasan, *Foreign Trade Regimes and Economic Development: India*
Koilpillai J. Charles

879 Willard W. Cochrane and Mary E. Ryan, *American Farm Policy, 1948–1973*
Don Paarlberg

882 Kenneth W. Dam, *Oil Resources: Who Gets What How?*
Colin Robinson

885 John F. Kain and John M. Quigley, *Housing Markets and Racial Discrimination*
Hugh O. Nourse

888 William G. Tyler, *Manufactured Export Expansion and Industrialization in Brazil*
José L. Carvalho

Journal of Political Economy

Volume 85 Number 5 October 1977

Articles

891 The Gibson Paradox and Historical Movements in Real Interest Rates
Robert J. Shiller and Jeremy J. Siegel

909 The Relevance of the Two-Sector Production Model [illegible] Theory
Ronald W. Jones and José A. Scheinkman

937 Trade Using One Commodity as a Means of Payment
Lloyd Shapley and Martin Shubik

969 Organized Futures Markets: Costs and Benefits
Lester G. Telser and Harlow N. Higinbotham

1001 Public Property and the Effects of Migration upon Other Residents of the Migrants' Countries of Origin and Destination
Dan Usher

1021 The Monetary Appreciation of Paintings
John Picard Stein

1037 Empirical Study of Scale Economies and Production Complementarity: The Case of Journal Publication
William J. Baumol and Yale M. Braunstein

1049 The IS-LM Model of Macroeconomic Equilibrium and the Monetarist Controversy
Neil Bruce

1063 Price Discrimination and Vertical Control: A Note
J. R. Gould

Comments

1073 Pollak and Wachter on the Household Production Function Approach
William A. Barnett

1083 Reply: "Pollak and Wachter on the Household Production Approach"
Robert A. Pollak and Michael L. Wachter

Book Reviews

1087 Robert J. Barro and Herschel I. Grossman, *Money, Employment and Inflation*
Benjamin M. Friedman

1092 W. M. Corden and Gerhard Fels, eds., *Public Assistance to Industry: Protection and Subsidies in Britain and Germany*
Edward Tower

1094 Richard B. Freeman, *The Overeducated American*
Cotton M. Lindsay

1096 Stephen Herbert Hymer, *The International Operations of National Firms: A Study of Direct Foreign Investment*
Thomas G. Parry

Journal of Political Economy

EDITED BY JACOB A. FRENKEL, SAM PELTZMAN,
AND GEORGE J. STIGLER
IN CO-OPERATION WITH OTHER MEMBERS OF THE
DEPARTMENT OF ECONOMICS AND THE
GRADUATE SCHOOL OF BUSINESS OF
THE UNIVERSITY OF CHICAGO
AND OUTSIDE REFEREES

Volume 85

February–December 1977

THE UNIVERSITY OF CHICAGO PRESS
CHICAGO, ILLINOIS

© 1977 BY THE UNIVERSITY OF CHICAGO. ALL RIGHTS RESERVED.
PUBLISHED FEBRUARY,
APRIL, JUNE, AUGUST, OCTOBER,
DECEMBER 1977
BY THE UNIVERSITY OF CHICAGO PRESS

Journal of Political Economy

Volume 85 December 1977

Articles

1099 Leisure, Home Production, and Work—the Theory of the Allocation of Time Revisited
Reuben Gronau

1125 The Principle of Unanimity and Voluntary Consent in Social Choice
Vernon L. Smith

1141 An Economic Analysis of Marital Instability
Gary S. Becker, Elisabeth M. Landes, and Robert T. Michael

1189 Location of Firms within a Square Central Business District
Donald J. O'Hara

1209 Transaction Costs and Interest Arbitrage: Tranquil versus Turbulent Periods
Jacob A. Frenkel and Richard M. Levich

1227 Money Wages, Prices, and Causality
Y. P. Mehra

1245 Exchange Rate Dynamics with Stock/Flow Interaction
Jürg Niehans

1259 Income Distributions in Two Experimental Economies
Raymond C. Battalio, John H. Kagel, and Morgan O. Reynolds

1273 Health, Investment in Health, and Occupational Choice
M. L. Cropper

1295 Inflationary Finance and Growth
Bijan B. Aghevli

1309 Erratum

Book Reviews

1311 A. K. Dixit, *The Theory of Equilibrium Growth*
A. Asimakopulos

1312 J. E. Meade, *The Just Economy*
Neil Bruce

1314 Carl H. Stem, John M. Makin, and Dennis E. Logue, eds., *Eurocurrencies and the International Monetary System*
Hans Genberg

1317 David Weatherall, *David Ricardo: A Biography*
Frank Whitson Fetter

1319 Index to Volume 85

In Memoriam

FABIAN BACHRACH

Harry G. Johnson

(May 26, 1923–May 8, 1977)

The editors record with deep regret the passing away of Harry G. Johnson, editor of the *Journal of Political Economy* 1961–1966 and 1970–1977.

Financial Flow Variables and the Short-Run Determination of Long-Term Interest Rates

---

Benjamin M. Friedman

Harvard University

Because transactions costs are smaller for allocating new cash flows than for reallocating existing asset holdings, financial flow variables are important determinants of investors' short-run asset demands. The demand-for-bonds equations implied by the resulting "optimal marginal adjustment" model of portfolio behavior constitute the demand side of a structural supply-demand model of the determination of the long-term interest rate. Empirical results, based on demand-for-bonds equations estimated using U.S. data for six major categories of bond market investors, support the optimal marginal adjustment model and show that the associated structural model of interest rate determination, which is restricted by the underlying demand-for-bonds equations, fits the data about as well as do previously developed unrestricted reduced-form term-structure equations.

The analytical frameworks which economists use to investigate various economic phenomena often bear little resemblance to market participants' conceptions of the processes in which they play some role. A striking example of this dichotomy is the question of how long-term interest rates are determined. Financial market participants, keenly sensitive to the fact that the immediate determination of bond yields takes place in a market in which securities are bought and sold, typically believe that the interplay between borrowers' issues of new bonds and investors' newly available cash flows represents an important determinant

The author is grateful to Kenneth Froewiss and Vance Roley for research assistance and to Gary Chamberlain, James Duesenberry, Patric Hendershott, Kenneth Froewiss, John Lintner, Franco Modigliani, Vance Roley, Robert Shiller, Allen Sinai, Gary Smith, and an anonymous referee for helpful comments on previous drafts of this paper. The author also gratefully acknowledges the National Science Foundation's support of the main part of this research, under grant SOC74-21027, and the National Bureau of Economic Research's support of computation done on its TROLL system.

[*Journal of Political Economy*, 1977, vol. 85, no. 4]

© 1977 by The University of Chicago. All rights reserved.

of long-term yields. In contrast, most economists have placed almost exclusive emphasis on the role in portfolio selection of relative asset returns and risks, which lead to reduced-form equations incorporating term-structure and price-expectations elements and have argued that, if quantity variables are relevant at all, it is not flows but stocks which matter.[1]

The object of this paper is to set forth and estimate, using U.S. data, a framework which reconciles these two conflicting views about the importance, or lack of importance, of financial flow variables in the determination of long-term interest rates. In particular, in the framework presented below, flow variables are important determinants of the short-run variation of long-term yields, while in the long run only stock variables and relative return and risk variables matter.

The vehicle which this paper uses to explore this question is a structural model of the determination of long-term interest rates, as previously described in some detail in Friedman (1974). The primary thrust of this model is to eschew the familiar unrestricted reduced-form term-structure equation and to use instead a set of structural equations representing supplies of and demands for long-term bonds. Since each such supply or demand equation is a function of the long-term bond yield, among other variables, the addition of a market-clearing identity equating total net excess supply (demand) to zero facilitates solving the model for the long-term bond yield itself. The principal usefulness of this structural model, for the purpose of this paper, is that it provides a way of focusing directly on those aspects of investors' behavior which lead to the short-run importance of financial flow variables. The form of the model presented in this paper is consistent with this orientation, in that it develops the role of the flow variables in detail but does not incorporate either complex lag structures on yield variables or a wide variety of exogenous variables other than straightforward asset return variables; such extensions remain for further research.[2]

Section I develops the basic framework of portfolio selection and portfolio adjustment, an "optimal marginal adjustment" model, within which financial flow variables play an important role in the determination of the short-run behavior of investors' asset demands. Section II briefly reviews the basic features of the overall structural model of long-term interest rate determination as used in this paper. Section III discusses some operational questions of estimation procedure and data. Section IV presents results for equations representing the demand for long-term bonds by the six major categories of investors in the U.S. corporate bond market, commenting in particular on the implications of the estimation results for the role of financial flow variables in influencing portfolio behavior. Section V presents two sets of *full-system dynamic simulation* results: for the seven-equation structural model consisting of the six estimated demand-for-bonds equations and a market-clearing identity (with bond supply taken as exogenous), and for the analogous nine-equation model incorporating the two supply-of-bonds equations developed in Friedman (1976). Section VI briefly summarizes the main conclusions of the paper.

---

[1] Culbertson's (1957) "market segmentation" views are something of an exception in their adaptability to the association of causal influence with flow variables. The work by Modigliani and Sutch (1966, 1967) and Modigliani and Shiller (1973) bears the label "preferred habitat" theory and therefore seems at first glance to allow for the role of flow variables, but no such variables appear in the empirical work. Silber (1970), Hendershott (1971), Bosworth and Duesenberry (1973), and Hendershott and Lemmon (1973) have introduced flow variables into their empirical models, but they have done so without providing an explicitly specified analytical framework to show why these variables are present.

[2] See Friedman (1974) for a discussion of lag structures and of exogenous variables which may play a role in determining investors' demands for bonds. Friedman (1976) uses such lag structures in empirical work on the supply of bonds by corporations.

I. The "Optimal Marginal Adjustment" Model of Portfolio Behavior

In a world in which transactions costs are nontrivial, it is useful to represent investors' portfolio behavior by a model which determines the desired long-run equilibrium portfolio allocation together with a model which determines the short-run adjustment toward the equilibrium allocation.

A familiar model of the selection of desired portfolio allocation, for a given investor or group of investors, is the linear homogeneous form[3]

$$\frac{A^*_{it}}{W_t} = \sum_k^N \beta_{ik} r_{kt} + \sum_h^M \gamma_{ih} X_{ht} + \pi_i, \qquad i = 1, \ldots, N, \tag{1}$$

where

A^*_{it}, $i = 1, \ldots, N$ = the investor's desired equilibrium holding of the ith asset at time period t ($\sum_i A^*_{it} = W_t$);

W_t = the investor's total portfolio size (wealth) at time period t;

r_{kt}, $k = 1, \ldots, N$ = the expected holding-period yield on the kth asset at time period t;

X_{ht}, $h = 1, \ldots, M$ = the values at time period t of additional variables which influence the portfolio allocation;

---

[3] Asset stocks and wealth are measured on an end-of-period basis throughout. See, for example, de Leeuw (1965), Brainard and Tobin (1968), and Hendershott (1971). Linear homogeneous asset demand functions may be derived from, for example, maximization of a negative exponential utility function in portfolio rate of return, where asset returns are normally distributed.

and the β_{ik}, γ_{ih}, and π_i are fixed coefficients which satisfy $\sum_i \beta_{ik} = 0$ for all k, $\sum_i \gamma_{ih} = 0$ for all h, and $\sum_i \pi_i = 1$. The role of the wealth homogeneity constraint is to require that any shift in an asset's share in the desired equilibrium portfolio be due to movements either of relevant yields (r_k) or of other variables (X_h), rather than to overall growth of the total portfolio itself; particularly for the case of equations representing the behavior of categories of investors, this assumption seems appropriate.[4]

The analytical role of the expected holding-period yield variables, r_k, $k = 1, \ldots, N$, is straightforward, although devising ways of empirically representing these expectations is highly complicated when potential capital gains or losses are involved. The usual presumption is that investors' demand for any asset responds positively to the own yield on that asset ($\beta_{ik} > 0$, $k = i$) and negatively to the yields on alternative assets ($\beta_{ik} < 0$, $k \neq i$). Because of differing portfolio objectives in terms of return and risk, differing effective transactions costs, and differing institutional and regulatory constraints, the group of assets which compete for a place in the collective portfolio of any investor or group of investors need not coincide with the entire menu of assets available in the economy. In practice, therefore, it is necessary to determine what potential substitutabilities are actually relevant in each case, and portfolio selection equations for different investor categories need not include identical groups of yield variables.[5]

The variables X_h, $h = 1, \ldots, M$, in the portfolio selection model (1) represent influences, other than the r_k measures of expected holding-period yields, which cause investors to change their desired equilibrium portfolio allocations among the various available assets. Following the mean-variance model of portfolio selection, these additional variables represent influences on investors' perceptions of return (to the extent that measures r_k are incomplete) and risk over the relevant holding period. Some investors may respond to certain specific influences, such as the anticipated volatility of nominal returns, while other investors may respond to different influences, such as anticipated price inflation. As is the case for the yield variables, therefore, portfolio selection equations for different investor categories need not include identical groups of nonyield variables.

Given the desired equilibrium portfolio allocation indicated by model (1), the usual description of investor behavior involves a shift of asset holdings which eliminates some, but not all, of the discrepancy between holdings $A_{i,t-1}$ at the end of the previous period and the new desired holdings A^*_{it}. One familiar representation of the resulting portfolio adjustment process is the stock adjustment model[6]

---

[4] For a further discussion of the rationale behind the wealth homogeneity constraint, see Friedman (1956) and de Leeuw (1965). For an alternative view, see Goldfeld (1966, 1969).

[5] In principle, of course, a sufficiently extreme return on any asset could bring about substitutions which do not normally take place. Nevertheless, the possibility that some individuals may buy private-placement bonds at a 25 percent yield, for example, does not warrant including the private-placement yield in equations for individuals' asset demands if that yield has varied only between 5 percent and 15 percent during the relevant sample period.

$$\Delta A_{it} = \sum_{k}^{N} \theta_{ik}(A^*_{kt} - A_{k,t-1}), \qquad i = 1, \dots, N, \tag{2}$$

where A_{it} = the investor's actual holding of the ith asset at time period t $(\sum_i A_{it} = W_t)$, and the θ_{ik} are fixed coefficients of adjustment such that $0 \leq \theta_{ik} \leq 1$, $k = i$, and $\sum_i \theta_{ik} = 1$ for all k. An analogous model which involves explicit adjustment of the portfolio proportions is

$$\Delta \alpha_{it} = \sum_{k}^{N} \theta_{ik}(\alpha^*_{kt} - \alpha_{k,t-1}), \qquad i = 1, \dots, N, \tag{3}$$

where

$$\alpha_{it} \equiv \frac{A_{it}}{W_t}, \qquad \left(\sum_i \alpha_{it} = 1\right),$$

and the constraint applicable to the θ_{ik} is the weaker $\sum_i \theta_{ik} = \bar{\theta}$ for all k (which does not require $\bar{\theta} = 1$).

At least three conceptually distinct phenomena may account for investors' failure to adjust their portfolios fully and immediately to whatever equilibrium allocations are consistent with each period's new values of the relevant variables: transactions costs, expectation formation lags, and perception lags. The lags associated with forming expectations and with perceiving current market developments in large part give rise empirically to the use of distributed lags on market yields to represent the expected holding-period yield variables (r_k) in portfolio selection model (1).[7] In contrast, the lags associated with transactions costs are of key importance to this paper's concern with the role of financial flow variables.

Transactions costs in the U.S. long-term debt markets take the form either of direct pecuniary charges or more indirect costs. Direct pecuniary charges include such factors as bid/asked spreads for institutional investors in the professional dealer market for debt securities and, in addition, brokerage fees for individual investors. Indirect costs for institutional investors include increased in-house overhead expenses associated with greater trading activity and participation in specialized forms of transactions, while for individual investors the most important factor is probably the opportunity cost of whatever time trading absorbs. Gathering and processing information may be a major component of the cost of trading for both institutions and individuals. Transactions costs also differ among types of debt securities. The bid/asked spread in the Treasury bill market, for example, is smaller than the spread for a trade of comparable size in the long-term bond market. Transactions cost distinctions at various levels exist even within the overall classification of long-term corporate bonds, as new issue purchases differ from secondary market trading, and privately placed securities usually tend to be highly illiquid.[8] Indirect transactions costs, such as the difficulty of obtaining rapid execution of a trade, also vary considerably from one type of security to another.

---

[6] Anderson (1964), de Leeuw (1965), and Goldfeld (1966) are a few examples of the application of the stock adjustment model to portfolio behavior. As Smith (1975) has pointed out, it is necessary to interpret each θ_{ik} as the partial effect on A_i of a unit increase in A^*_k accompanied by a unit increase in W.

[7] See Friedman (1974) for further discussion of these two sources of lagged adjustments.

One implication of these transactions costs is that it is easier (cheaper) for most investors to allocate new cash flows—including both net wealth increments and repayments such as dividends, coupon interest, and maturities—than to reallocate current asset holdings.[9] As a result, the simple stock adjustment models (2) and (3) are inadequate in failing to distinguish clearly between new cash flows and previous-period wealth (including capital gains and net of repayments).

Figure 1 illustrates explicitly this overrestrictiveness of portfolio adjustment models (2) and (3) by representing portfolio adjustment model (3) for the case of an investor whose portfolio consists of two assets only, holdings of which are indicated by distances from the origin in the two-dimensional diagram. For fixed asset prices P_1 and P_2 for assets A_1 and A_2, respectively, the straight line from $[0, (W_{t-1}/P_2)]$ to $[(W_{t-1}/P_1), 0]$ represents the investor's wealth constraint in time period $(t - 1)$, and the parallel line from $[0, (W_t/P_2)]$ to $[(W_t/P_1), 0]$ represents the wealth constraint after an increment $\Delta W_t > 0$. Point Y_{t-1} indicates the investor's previous period asset holdings $(A_{1,t-1}, A_{2,t-1})$, and point Y_t indicates the holdings (A_{1t}, A_{2t}) which would result from no reallocation of the existing asset holdings and allocation of the wealth increment ΔW_t in proportion to the actual holdings at time period $(t - 1)$. Point Z_t^* indicates the desired equilibrium holdings (A_{1t}^*, A_{2t}^*). Point Z_t indicates the actual holdings (A_{1t}, A_{2t}) which result from portfolio adjustment model (3).

---

[8] In recent years, quoted bid/asked spreads in the professional corporate bond market have fallen to fairly low levels for a wide selection of securities as trading volume in these securities has grown. Nevertheless, since the market as a whole has remained quite thin, institutional investors are conscious of the extent to which sizable transactions that they may attempt can cause both bid and asked quotations to shift in the direction representing a market deterioration from their current point of view. Hence quoted bid/asked spreads understate true direct pecuniary transactions costs. With respect to private placements, trading liquidity is limited but is nevertheless often greater than Shapiro and Wolf (1972) implied.

[9] Duesenberry (1963) has made this point, as have a number of subsequent writers.

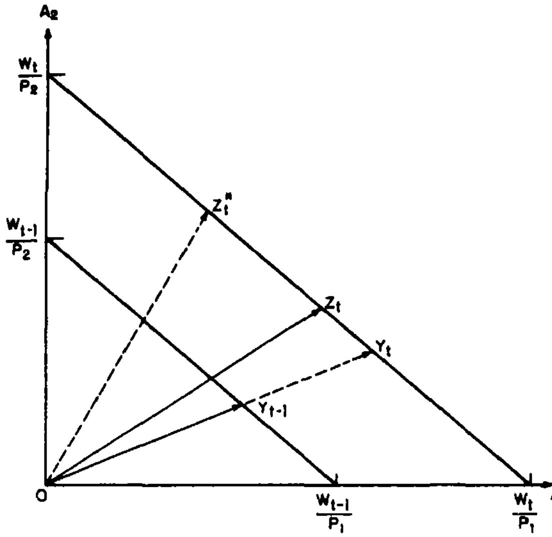


FIG. 1.—Portfolio adjustment model (3), two-asset case

The necessary and sufficient conditions for points Z_t and Z_t^* in figure 1 to be identical are $\theta_{11} - \theta_{12} = \theta_{22} - \theta_{21} = 1$. In the absence of such a complete adjustment model, Z_t and Z_t^* are not identical, and the model's symmetrical adjustment factor, applicable to both assets, is indicated by the ratio of the lengths of line segments Y_tZ_t and $Y_tZ_t^*$,

$$\frac{Y_tZ_t}{Y_tZ_t^*} = \frac{\alpha_{1t} - \alpha_{1,t-1}}{\alpha_{1t}^* - \alpha_{1,t-1}} = \frac{\alpha_{2t} - \alpha_{2,t-1}}{\alpha_{2t}^* - \alpha_{2,t-1}} = \theta_{11} + \theta_{22} - \bar{\theta}, \quad (4)$$

which is independent of the wealth increment ΔW_t.[10]

Hence the standard stock adjustment model—even in the full form including "cross-discrepancies"[11]—is too restrictive in that it does not

---

[10] As defined for equation (3), $\bar{\theta}$ is the identical sum of each column in the matrix of adjustment coefficients. The symmetry of the two-asset case, by which $(\alpha_{1t} - \alpha_{1,t-1})/(\alpha_{1t}^* - \alpha_{1,t-1}) = (\alpha_{2t} - \alpha_{2,t-1})/(\alpha_{2t}^* - \alpha_{2,t-1})$, does not hold in the general case; for three or more assets, the $(\alpha_{it} - \alpha_{i,t-1})/(\alpha_{it}^* - \alpha_{i,t-1})$ are not necessarily equal but are still independent of ΔW_t. The purpose of figure 1 is to illustrate this independence of the $(\alpha_{it} - \alpha_{i,t-1})/(\alpha_{it}^* - \alpha_{i,t-1})$ and ΔW_t in the stock adjustment model. The graphical convenience of two dimensions is the only reason for using a two-asset model in the figure.

[11] Brainard and Tobin (1968) have emphasized the importance of the dependence of the adjustment of any asset holding ΔA_{it} on the discrepancy $(A_{kt}^* - A_{k,t-1})$ for other assets $k \neq i$. An implication of this point for model (2) is that, if $\theta_{ik} = 0$ for $k \neq i$, so that the model is equivalent to

$$\Delta A_{it} = \theta_{ii}(A_{it}^* - A_{i,t-1}), \qquad i = 1, \ldots, N, \qquad (2')$$

then $\theta_{ii} = 1$ for all i, and the model requires full adjustment of each asset holding $A_{i,t-1}$ to the desired holding A_{it}^*, so that $A_{it} = A_{it}^*$ for all i. The analogous implication of this reflect the greater adjustability of new cash flows in comparison with existing asset holdings. In the context of a disaggregated (by investor) model, transactions costs should render each investor category's asset demands dependent in the short run on the investable cash flow of that particular category of investors.

In principle, the ideal way to represent the effect on portfolio behavior of the easier (less costly) allocation of cash flow would be a model incorporating some set of discontinuities or switch-over points. If $(A_{it}^* - A_{i,t-1}) \geq 0$ for all i, then no selling of previously held assets is necessary. If $(A_{it}^* - A_{i,t-1}) < 0$ for some i, then ΔW_t is insufficient to meet all desired purchases; ΔW_t should be applied only to purchasing assets A_i for which $(A_{it}^* - A_{i,t-1}) > 0$. The discontinuous aspects of such a model would render it empirically intractable, however, and the use of data at any level but that of the individual portfolio would be inappropriate in this context.

Perhaps the simplest way of incorporating explicit flow effects on portfolio adjustments without introducing such discontinuities is to separate the reallocation of existing holdings from the allocation of the new cash flow, applying fixed coefficients to the flow itself:[12]

$$\Delta A_{it} = \sum_{k}^{N} \theta_{ik}(\alpha_{kt}^* \cdot W_{t-1} - A_{k,t-1}) + \delta_i \Delta W_t, \quad i = 1, \ldots, N, \quad (5)$$

---

point for model (3) is that, if $\theta_{ik} = 0$ for $k \neq i$, so that the model is equivalent to

$$\Delta \alpha_{it} = \theta_{ii}(\alpha_{it}^* - \alpha_{i,t-1}), \quad i = 1, \ldots, N, \quad (3')$$

then the θ_{ii} are identical for all i, and the model requires equal fractional adjustment of each asset proportion $\alpha_{i,t-1}$ toward the desired proportion α_{it}^*.

[12] Model (5) is similar, but not identical, to Brainard and Tobin's (1968) rendering of model (2). Since $A_{kt}^* = \alpha_{kt}^* W_t$ for all k, the Brainard-Tobin equation (p. 107) is analogous to

$$\Delta A_i^* = \sum_{k}^{N} \theta_{ik}(A_{kt}^* - A_{k,t-1}) + \delta_i \Delta W_t, \quad i = 1, \ldots, N. \quad (5')$$

As Smith (1975) has pointed out, model (5′) "overdescribes" the portfolio disequilibrium, since $\sum_k (A_{kt}^* - A_{k,t-1}) = \Delta W_t$. To implement the Brainard-Tobin model, therefore, it is necessary either to eliminate the term in ΔW_t, thereby rendering (5′) identical to (2), or to delete the term in $(A_{kt}^* - A_{k,t-1})$ for some arbitrary k, as Brainard and Tobin did implicitly in their table 3 (p. 108). (A further difference between either [5] or [5′] and the Brainard-Tobin model is not fundamental: Brainard and Tobin further divided ΔW_t into capital gains [on equities] and all other wealth increments, i.e., they replaced $[\delta_i \Delta W_t]$ by $[\varepsilon_i \Delta W_{1t} + \eta_i \Delta W_{2t}]$, where ΔW_1 is capital gains and ΔW_2 is the sum of all other elements of the total wealth increment ΔW. For the capital gains term ΔW_{1t}, they assumed that the ε_i coefficient in the equity equation has value unity and all other ε_i have value zero. For the other wealth flow term ΔW_{2t}, they assumed that the η_i coefficient in the deposit equation has value unity and all other η_i have value zero.) Some researchers have based empirical work on model (5); see, for example, Kopcke (1973). For empirical work relying specifically on Brainard and Tobin (1968), see, for example, Smith and Brainard (1974). For an alternative viewpoint on empirical implementation of the Brainard-Tobin model, see Ando and Modigliani (1975).

where the δ_i are fixed coefficients such that $0 \leq \delta_i \leq 1$ for all i and $\sum_i \delta_i = 1$, and the constraint on the θ_{ik} is $\sum_i \theta_{ik} = \bar{\theta}$ for all k, as in model (3). The first term on the right-hand side of model (5) represents the reallocation of the existing wealth W_{t-1}, while the second term represents the allocation of the flow of new wealth ΔW_t.

Although model (5) may represent some improvement over models (2) and (3), in that the adjustments $(\alpha_{it} - \alpha_{i,t-1})/(\alpha^*_{it} - \alpha_{i,t-1})$ (analogous to ratio $[Y_tZ_t]/[Y_tZ^*_t]$ in fig. 1) depend explicitly upon ΔW_t, the fixed coefficients δ_i render it inadequate for most investors, except for extremely short time periods. In particular, assuming that investors always allocate new cash flows in fixed proportions, regardless of market yields and other variables relevant to portfolio behavior, is in some sense more restrictive than assuming no distinction at all between allocation of the new cash flow and reallocation of existing holdings. Furthermore, only by chance will a growing portfolio ever reach equilibrium if it is invested according to model (5), even if the desired proportional allocations α^*_{kt} are constant.

A compromise between these two restrictive approaches is a model which allocates the flow of new investable funds according to whatever proportions portfolio selection model (1) indicates are the desired equilibrium proportions for the portfolio as a whole—that is, a model which, away from equilibrium, uses equilibrium total proportions to allocate at the margin. The "optimal marginal adjustment" model indicated by this specification is an analogue of model (5) which replaces that model's invariant flow allocation coefficient δ_i for each asset by the corresponding current desired proportion α^*_{it}:[13]

$$\Delta A_{it} = \sum_{k}^{N} \theta_{ik}(\alpha^*_{kt} \cdot W_{t-1} - A_{k,t-1}) + \alpha^*_{it}\Delta W_t, \qquad i = 1, \ldots, N. \qquad (6)$$

Such a model seems especially appropriate for investors participating in private securities markets, in which some level of fairly continuous activity maintains the flow of information about new investment opportunities. Institutional investors such as insurance companies and pension funds, for example—which participate not only in the market for publicly offered debt issues but also in the markets for privately placed corporate debt securities, commercial mortgages, and other forms of negotiated

---

[13] Modigliani (1972) used a model which is similar to the optimal marginal adjustment model (6) but which allocates the current-period flow according to the observed previous-period stock allocation $\alpha_{i,t-1}$. The optimal marginal adjustment model allocates the flow according to the desired equilibrium stock allocation α^*_{it}. Although the two models appear similar, the rationale is very different. The optimal marginal adjustment model is a description especially relevant for institutional investors operating in markets with transactions (including information) costs. Modigliani's model, which he applied in particular to household demand for savings deposits, explicitly postulates a short-run desired allocation which differs from the (long-run) equilibrium desired allocation.

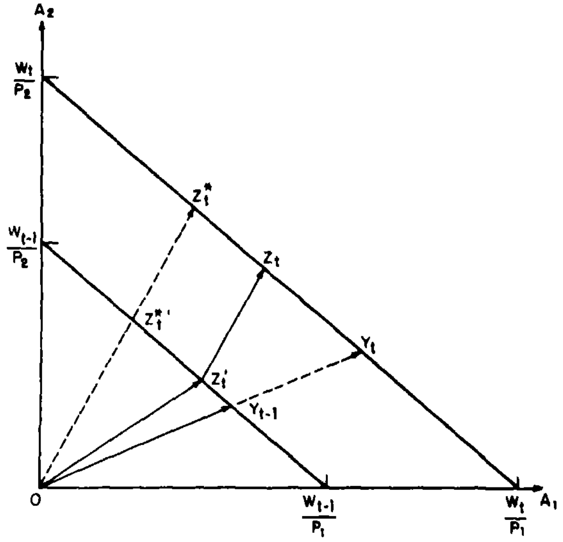


FIG. 2.—Portfolio adjustment model (6), two-asset case

loans- are typically reluctant to undertake sharp negative swings in their new purchasing activities for long periods of time. The optimal marginal adjustment model also serves to approximate this type of smoothed behavior.

Figure 2 illustrates the advantage of the optimal marginal adjustment model (6) in the two-asset case analogous to that depicted in figure 1. Points Y_{t-1}, Y_t, and Z_t^* are as in figure 1. The first term on the right-hand side of model (6) represents the reallocation of the existing wealth W_{t-1}.[14] Figure 2 depicts this reallocation as a move along the W_{t-1} wealth constraint from point Y_{t-1} toward point $Z_t^{*\prime}$; point $Z_t^{*\prime}$ is the intersection of the W_{t-1} wealth constraint and the ray from the origin which passes through point Z_t^*. Except for the special case of full adjustment ($\theta_{ii} = 1$ for all i, and $\theta_{ik} = 0$ for $k \neq i$), this reallocation process leads not to point $Z_t^{*\prime}$ itself but rather to point Z_t' which lies between $Z_t^{*\prime}$ and Y_{t-1} on the W_{t-1} wealth constraint. The portfolio adjustment model's θ_{ik} coefficients determine the adjustment ratio $(\overline{Y_{t-1}Z_t'})/(\overline{Y_{t-1}Z_t^{*\prime}})$. If wealth is constant, the movement to point Z_t' is the only component of the portfolio adjustment, and model (6) reduces to model (3).

---

[14] The W_{t-1} variable in model (6) includes capital gains but excludes repayments, which become part of the gross cash flow variable ΔW_t. For simplicity, the example shown in figure 2 assumes no repayments. As in figure 1, it also abstracts from capital gains by assuming constant asset prices.

The second term on the right-hand side of model (6) represents the allocation of the wealth increment ΔW_t according to the current desired proportions α^*_{it}. Figure 2 depicts this allocation as a movement from point Z'_t on the W_{t-1} wealth constraint to point Z_t on the W_t wealth constraint, along a line parallel to the ray from the origin which passes through point Z^*_t. Point Z_t indicates the new holdings (A_{1t}, A_{2t}) which result from portfolio adjustment model (6). In contrast to model (3) as illustrated in figure 1, figure 2 shows that, for given W_{t-1}, the symmetrical adjustment ratio $(Y_tZ_t)/(Y_tZ^*_t) = (\alpha_{1t} - \alpha_{1,t-1})/(\alpha^*_{1t} - \alpha_{1,t-1}) = (\alpha_{2t} - \alpha_{2,t-1})/(\alpha^*_{2t} - \alpha_{2,t-1})$ in model (6) varies positively with the flow ΔW_t.[15]

Hence the optimal marginal adjustment model, without sacrificing either tractability or suitability for empirical application to aggregative data, serves the intended purpose of relaxing the restrictiveness of the standard stock adjustment model with respect to investors' allocation of their cash flows. The model therefore leads to asset demand equations in which investable cash flow variables matter for short-run allocations. In particular, the greater the new cash flow, the faster the portfolio's overall adjustment toward equilibrium allocation. Section II shows how a structural model relying on asset demand equations of this form in turn renders flow variables determinants of the short-run variation of asset yields.

II. The Market-clearing Structural Model

Economists' models of the determination of long-term interest rates have usually adopted a term-structure approach, according to which the value of a representative long-term rate of nominal interest depends upon the value of a representative short-term rate of nominal interest, with the spread between the two rates determined by some set of economic factors which influence the behavior of borrowers or lenders or both. Among the familiar factors which various writers have suggested including as determinants of this spread are the expected future path of the short-term rate of interest (usually represented by some distributed lag on past values of the short-term rate; see, e.g., Meiselman [1962], Malkiel [1966], and Modigliani and Sutch [1966]), a premium reflecting the differential liquidity of short- and long-term debt instruments (Hicks [1939]; Kessel [1971]), expected future price inflation in product markets (usually represented by some distributed lag on past values of a chosen price inflation index; Modigliani and Shiller [1973]), and an index of current central bank monetary policy beyond interest rates themselves (e.g., the recent rate of growth of some monetary or reserve aggregate).[16]

---

[15] As in figure 1, the symmetry of adjustments is a particular result due to the two-asset model. The dependence of the $(\alpha_{it} - \alpha_{i,t-1})/(\alpha^*_{it} - \alpha_{i,t-1})$ ratios on ΔW_t in model (6) holds for any number of assets. See n. 10 above.

The prevailing empirical methodology of the term-structure approach has been a model consisting of a single *unrestricted* reduced-form equation with the nominal long-term interest rate as the dependent variable. The use of this methodology implies an acceptance of two key assumptions, at least as close approximations. First, it assumes that not only the short-term interest rate but also the determinants of the long/short spread are exogenous with respect to the actions of participants in the market for long-term debt securities.[17] Second—and more important, from the standpoint of this paper--it assumes that the way in which participants in the market for long-term securities, either individually or in the aggregate, adjust their actions in that market in response to any or all of the determinants of portfolio behavior does not matter for the ex post outcome for the long-term interest rate. In particular, this second assumption implies that the quantities of long-term securities bought or sold, either by individual transactors or for the market in aggregate, do not influence the ex post outcome for the long-term interest rate. A few writers have suggested relaxing this second assumption somewhat by incorporating exogenous supplies of long-term securities as a further determinant of the long/short spread, but they have done so within the familiar unrestricted reduced-form equation methodology of the term-structure approach.[18]

The structural model of long-term interest rate determination, as applied in this paper, drops the second of these two assumptions and specifies equations directly representing the portfolio behavior of bond market participants. The addition of a market-clearing constraint, equating the sum of the demands for long-term debt securities to the sum of the supplies of long-term debt securities, enables the structural model to determine the long-term interest rate (i.e., the own yield) which appears as a right-hand-side variable in each structural demand or supply equation.[19] The resulting implicit expression for the long-term interest

[16] See, for example, Feldstein and Eckstein (1970). Their work is within the usual approach empirically, in that they estimated an unrestricted equation with the long-term interest rate as the dependent variable. Conceptually, however, they rejected the term-structure approach, omitting the short-term yield altogether and interpreting the estimated equation as an inverted form of money-demand function.

[17] See Feldstein and Chamberlain (1973) for an unusually careful treatment of this issue, within the context of the single-equation reduced-form approach.

[18] Feldstein and Eckstein (1970) used the stock supply of U.S. government securities in this way. Fair and Malkiel (1971) used "visible" flow supplies of new corporate bonds in reduced-form equations for spreads between yields on different types of long-term securities; the determination of long-term yield spreads in this way is a corollary to the determination of the long/short spread in the term-structure approach.

[19] Silber (1970) and Hendershott and Lemmon (1973) have also applied this approach. Bosworth and Duesenberry (1973) used this approach conceptually, but not in their

rate is a kind of reduced-form equation (it may be nonlinear) which is *restricted* by the associated structural supply and demand equations.

Hence the key methodological difference between the structural approach and the familiar term-structure approach to modeling long-term interest rate determination is essentially equivalent to the distinction between *restricted* and *unrestricted* estimation. The principal advantage of the structural approach is its ability to use the underlying theory of portfolio behavior to constrain the implied equation for the long-term interest rate, which in turn facilitates directly testing hypotheses about portfolio behavior. The structural approach imposes upon the researcher the discipline of having to acknowledge explicitly that any factor hypothesized to influence the bond yield can do so only by influencing some issuer's supply of bonds or some investor's demand for bonds (or both). To the extent that less-than-infinite elasticities of substitution create "preferred habitats" which render quantity variables relevant, for example, or to the extent that less-than-infinite adjustment speeds render quantity flow variables relevant as well as quantity stock variables, in the structural model such factors can influence the determination of the long-term interest rate only by influencing the behavior of borrowers and lenders.

Since the long-term interest rate is clearly a jointly determined variable in this model, along with the demand and supply variables, it is necessary to use estimation techniques which avoid inconsistencies to which ordinary least-squares procedures would be subject because of the nature of the model's simultaneity. The empirical work described in Section IV below focuses entirely on the investor side of the corporate bond market, while Friedman (1976) develops an analogous treatment of bond issuers. The simulation results presented in Section V are derived from a simulation of the bond demand model alone, with bond supply taken as exogenous, and also from a separate simulation of the combined model of bond demand and bond supply.

The particular long-term interest rate to be determined in this model is the observed new-issue yield on long-term bonds issued by utility companies rated Aa by Moody's Investors Service, Inc. Aa-rated utility bonds provide the greatest continuity, in terms of the frequency of new issues; they are also most representative of new-issue activity in the U.S. corporate bond market. Previous studies of long-term interest rate determination using the term-structure approach have relied on indexes of yields either on new issues or seasoned issues, but the new-issue yield is likely to be superior for several reasons. First, trading in the corporate bond market involves either new issues or recent issues to a far greater

---

empirical work; instead, for each asset, they typically dropped one sector's demand equation and estimated an unrestricted equation with the asset's yield as the dependent variable and the omitted sector's demand as an independent variable.

extent than seasoned issues, and quoted price movements among seasoned issues are often just a reflection of what is happening in the new-issue market. Second, because of thin trading markets, problems of measurement are considerably smaller for new issues than for seasoned issues. Third, Modigliani and Shiller (1974) have shown that differences in coupon rates between current new issues and the issues used in constructing seasoned yield indexes introduce a form of bias into the seasoned yield index itself.

The structural equations presented in Section IV represent investors' demands for long-term corporate bonds. These equations follow the optimal marginal adjustment model of portfolio behavior developed in Section I in their treatment of financial flow variables. In addition, by disaggregating demand according to six major categories of corporate bond investors, these equations permit the various investing sectors' cash flows to exert differential effects on the determination of the long-term yield. The six categories of investors treated as endogenous in this model are (1) life insurance companies, (2) other insurance companies, (3) private pension funds, (4) state and local government retirement funds, (5) mutual savings banks, and (6) households.[20] As of the end of 1974, as table 1 shows, these investors held approximately 94 percent of all outstanding long-term corporate bonds issued in the United States. The model treats as exogenous all other investors' demands for long-term corporate bonds.

The portfolio behavior model developed in Section I deals with a single investor's, or investor category's, demands for all assets (or liabilities) in the markets for which the investor participates. By contrast, the structural model used in this paper focuses on one asset (long-term corporate bonds) and includes behavioral equations describing the demand for that single asset by each major category of investors participating in that particular market. Hence each of the six structural demand equations is implicitly an element of a set of demand equations which satisfy the "adding-up" constraints specified in Section I.

As Ladenson (1971) and Smith (1975) have shown, however, it is not necessary to use constrained estimation techniques to guarantee that the parameter estimates of the full set of demand equations for any investor category satisfy the "adding-up" constraints, and so there is no inconsistency involved in estimating only one demand equation, rather than the entire set, for each category of investors. In principle, however, a full model including all investors and all markets (i.e., all assets) would be preferable for two reasons. First, it would render all yields—not just the own-yield on the one asset under consideration—jointly determined

---

[20] The household sector consists primarily of individuals but also includes nonprofit organizations and bank-managed personal trusts.

TABLE 1

DISTRIBUTION OF U.S. CORPORATE BOND HOLDINGS

Bonds Held By:	Amount ($ Billions)	Percentage
Life insurance companies*	97.9	34.1
State and local government retirement funds*	57.8	20.1
Households*	54.5	19.0
Private pension funds*	34.5	12.0
Mutual savings banks*	14.1	4.9
Other insurance companies*	11.0	3.8
Commercial banks	7.3	2.5
Mutual funds	3.8	1.3
Foreign investors	3.5	1.2
Brokers and dealers	2.7	0.9
Total	287.1	100.0

NOTE.—Detail may not add to total because of rounding; all data are at par value except foreign; groups marked by asterisk are endogenous in model.

variables. Second, it would permit the researcher to adopt the philosophy as well as the mechanics of Brainard and Tobin (1968) by examining the implications for other asset demand equations of the presence of a given variable in any one asset demand equation. Nevertheless, the construction of such a full model lies beyond the scope of this paper.[21]

III. Estimation Procedures and Data

Before proceeding to the empirical results in Section IV, it is worthwhile to comment briefly on both the estimation procedures and the data used.

Estimation Procedures

The long-term corporate bond demand equation estimated for each of the six categories of investors combines the linear homogeneous selection model and the optimal marginal adjustment model, as developed in Section I.[22] To see the implications of the combination of selection model

---

[21] For examples of efforts toward building such a full model, see Bosworth and Duesenberry (1973), Hendershott and Lemmon (1973), Smith and Brainard (1974), Ando and Klein (1975), and Fromm and Sinai (1975).

[22] Especially for those categories of investors which are not heavily involved in private placement markets, preliminary experiments also used an alternative portfolio adjustment model of the form

$$\Delta\alpha_{it} = \sum_{k}^{N} \left(\theta_{ik} + \phi_{ik}\frac{\Delta W_t}{W_{t-1}}\right)(\alpha^*_{kt} - \alpha_{k,t-1}), \qquad i = 1, \ldots, N, \tag{6'}$$

where the ϕ_{ik} coefficients satisfy the same constraint which applies to the θ_{ik} coefficients in model (3). This adjustment model, which also allows the flow ΔW_t to influence the portfolio allocation but does so in a more general way than model (6), uniformly led to less satisfactory empirical results than those obtained using the optimal marginal adjustment model (6).

(1) and adjustment model (6) into the overall model of portfolio behavior, it is useful to expand the resulting equation, grouping the variables in the following way:

$$\begin{aligned}\Delta A_{it} = {} & \pi_i \cdot \Delta W_t + \left[\sum_k (\pi_k \cdot \theta_{ik})\right] \cdot W_{t-1} \\ & + \beta_{ii} \cdot r_{it} \cdot \Delta W_t + \left[\sum_k (\beta_{ki} \cdot \theta_{ik})\right] \cdot r_{it} \cdot W_{t-1} \\ & + \sum_{k \neq i} \left\{\beta_{ik} \cdot r_{kt} \cdot \Delta W_t + \left[\sum_j (\beta_{jk} \cdot \theta_{ij})\right] \cdot r_{kt} \cdot W_{t-1}\right\} \qquad (7) \\ & + \sum_h \left\{\gamma_{ih} \cdot X_h \cdot \Delta W_t + \left[\sum_k (\gamma_{kh} \cdot \theta_{ik})\right] \cdot X_h \cdot W_{t-1}\right\} \\ & - \theta_{ii} \cdot A_{i,t-1} - \sum_{k \neq i} (\theta_{ik} \cdot A_{k,t-1}).\end{aligned}$$

The flow ΔW_t and lagged wealth stock W_{t-1} enter equation (7) linearly, but in neither case is the coefficient sign known a priori; π_i is simply the constant term corresponding to asset A_i in selection model (1), and π_k, $k \neq i$, are the analogous constant terms in the selection model equations for other assets A_k, $k \neq i$.[23] The own yield r_{it} enters equation (7) twice, in nonlinear form both times. The coefficient of the $(r_{it} \cdot \Delta W_t)$ term is $\beta_{ii} > 0$, but the coefficient of the $(r_{it} \cdot W_{t-1})$ term is a sum of coefficients which is of unknown sign a priori.[24] The yields on alternative assets r_{kt}, $k \neq i$, also enter equation (7) twice, in nonlinear form both times. Once again, the coefficient of each $(r_{kt} \cdot \Delta W_t)$ term, for $k \neq i$, is $\beta_{ik} < 0$, but the coefficient of each $(r_{kt} \cdot W_{t-1})$ term is a sum of coefficients which is of unknown sign a priori. Each nonyield variable X_h also enters in a precisely analogous way. Finally, the lagged own stock $A_{i,t-1}$ enters linearly with coefficient $-\theta_{ii} < 0$, and the lagged stocks of competing assets $A_{k,t-1}$, $k \neq i$, enter linearly with coefficients of unknown sign.[25]

---

[23] This feature of the model is in sharp contrast to the work of Bosworth and Duesenberry (1973) and Hendershott and Lemmon (1973) who, without explicitly developing the underlying portfolio adjustment model, emphasized the linear appearance of cash flow variables. A consequence of the optimal marginal adjustment model is that it is the *nonlinear terms*, with cash flow multiplied by another variable, which matter. The cash flow or lagged wealth stock in linear form could just as well have zero coefficient, as many do in the estimation results presented in Section IV.

[24] The indeterminancy of $\sum_k (\beta_{ki} \cdot \theta_{ik})$ is due to the unknown signs of θ_{ik}, $k \neq i$, since $\beta_{ii} > 0$ and presumably $\beta_{ik} < 0$, $k \neq i$. Although an intuitive guess is $\theta_{ik} < 0$, $k \neq i$ (which would then render the sum unambiguously positive), there is no reason to presume $\theta_{ik} < 0$, $k \neq i$, since $\alpha^*_{it} \lesseqgtr \alpha_{i,t-1}$ itself implies $\alpha^*_{kt} \gtreqless \alpha_{k,t-1}$ for at least some $k \neq i$. Hence the adjustment coefficients θ_{ik}, $k \neq i$, are of unknown sign. I am grateful to William Brainard, Gary Smith, and James Tobin for an illuminating discussion of this point.

[25] See n. 24 above.

TABLE 2

PRINCIPAL ASSETS FOR SIX INVESTOR CATEGORIES

Category and Assets	Percentage of Total Financial Assets
Life insurance companies:	
Corporate bonds	38.3
Commercial mortgages	25.0
Equities	8.7
Residential mortgages	8.8
Subtotal	80.8
Other insurance companies:	
Corporate bonds	16.1
Municipal bonds	47.1
Equities	18.3
U.S. government securities	5.4
Subtotal	86.9
Private pension funds:	
Corporate bonds	29.1
Equities	55.6
U.S. government securities	4.6
Subtotal	89.3
State and local government retirement funds:	
Corporate bonds	61.6
Equities	23.5
U.S. government securities	5.6
Subtotal	90.7
Mutual savings banks:	
Corporate bonds	12.8
Residential mortgages	40.7
Commercial mortgages	27.7
U.S. government securities	6.3
Subtotal	87.5
Households:	
Corporate bonds	2.5
Liquid assets	45.4
Demand deposits and currency	8.0
Time deposits and savings accounts	31.9
U.S. government securities	5.5
Equities	24.0
Municipal bonds	2.8
Subtotal	74.7

Most of the six investor categories examined here hold at least some amounts of a large number of different types of assets. For each category, the procedure used in estimation was to select the few assets which, along with corporate bonds, comprise the major elements in the investors' aggregate portfolios. Table 2 lists such likely asset groups for each of the six investor categories and shows the percentage of the sector's total financial portfolio for which each asset accounted as of year-end 1974. As the discussion in Section I notes,[26] equation (7) is not identifiable in the form including not only W_{t-1} and $A_{i,t-1}$ but also $A_{k,t-1}$ for all $k \neq i$. The procedure used to deal with this problem was to estimate each equation once including W_{t-1} and $A_{i,t-1}$ only (thereby implicitly imposing the constraint that the cross-adjustment coefficients θ_{ik} are identical for all $k \neq i$); once including $A_{i,t-1}$ and all $A_{k,t-1}$, $k \neq i$, but excluding W_{t-1}; and then including W_{t-1}, $A_{i,t-1}$ and $A_{k,t-1}$ for only some subset of $k \neq i$. Final selection among these alternative constrained subspecifications depended on goodness of fit and statistical properties. With respect to the yields, the analogous procedure used was to include only some subset of yields r_{kt} on alternative assets $k \neq i$.

---

[26] See especially n. 12, above.

As the discussion in Section II emphasizes, it is necessary to use some form of consistent estimation procedure to avoid simultaneity problems due to the joint determination of the own yield on long-term corporate bonds and the demand variables which are the dependent variables in the six separate equations of form (7). As is typically the case in multi-equation models, it is impossible to apply the two-stage least-squares method directly because there are too many exogenous variables to permit ordinary least-squares estimation of the system's reduced form, as this method requires. The procedure used here follows Brundy and Jorgenson (1971) in using as instrumental variables not only the leading principal components of the full-system set of exogenous variables but also, on an equation-by-equation basis, the single-equation sets of exogenous variables themselves.

A final issue in estimation procedure is the question of an intercept term. Equation (7) has no such term, but it is probably accurate to consider the portfolio behavior model developed in Section I as a linear approximation to a more complex behavioral pattern, and an intercept may follow from linearization. The procedure used was to include or exclude an intercept in each demand equation according to the t-statistic.

Data

The primary data source for the stock and flow quantities used in this model is the Board of Governors of the Federal Reserve System's flow-of-funds accounts (1973, and subsequently). The sample period consists of 56 quarterly observations beginning in 1960:I and ending in 1973:IV.

The bond demand variables—ΔA_{it} in equation (7)—are the net purchases of corporate bonds during the quarter, seasonally adjusted, for each of the six investor categories.

The wealth flow variables—ΔW_t in equation (7)—are the net acquisitions of financial assets, seasonally adjusted, for four investor categories: non-life insurance companies, private pension funds, state and local government retirement funds, and mutual savings banks. For life insurance companies, ΔW_t is the net acquisition of financial assets, seasonally adjusted, less net policy loans; the reason for the subtraction is that policy loans are exogenous to the portfolio behavior of life insurance companies, which have available for investment only that portion of their cash flow which is left after policy loans. For households, ΔW_t is the net acquisition of financial assets, seasonally adjusted, plus the capital gains portions of pension fund and life insurance reserves.

Since the "net acquisitions of financial assets" concept excludes capital gains, defining the six ΔW_t variables on this basis effectively imposes on the portfolio model a part of the Brainard-Tobin assumption discussed in connection with model (5) in Section I. In particular, it assumes that investors do not respond within the current quarter to shifts of their portfolios due to capital gains.[27] It is also important to point out that the "net acquisitions of financial assets" concept differs from the ΔW_t flow variable discussed in Section I precisely because it is "net." As the discussion in Section I explains, the relevant flow concept should include repayments. Nevertheless, gross cash flow data are simply unavailable for most categories of investors.

The asset stock variables A_{kt}, for all six investor categories, are constructed by decrementing backward from the end-of-year stocks for 1973, using seasonally adjusted quarterly flows.[28] For those assets for which the flow-of-funds data incorporate market valuation changes, however, the procedure used was to separate the total quarterly flows into net purchases and market valuation components and to use the latter without seasonal adjustment.[29] The total wealth variables W_t are constructed analogously to the asset stock variables for all six sectors.

IV. Estimation Results

The equations listed below are the results of estimating equation (7) for each of the six major categories of corporate bond investors. The dependent variable is in each case denominated in millions of dollars. The numbers in parentheses are ratios of estimated values to standard errors for each coefficient; because of the estimation procedure used, these ratios are asymptotically distributed as t-statistics but are not necessarily distributed as t-statistics in small samples. The $\bar{R}^2$ is the coefficient of determination adjusted for degrees of freedom, SE is the standard error of estimate, and D-W is the Durbin-Watson statistic. The variable symbols listed below are consistent for all six equations, with letter superscripts indicating distinctions among corresponding variables for different investor categories. An asterisk superscript indicates terms for which the equation is estimated using fitted values of the variable from the first stage of the instrumental variables procedure.[30]

---

[27] See nn. 12 and 14; see also n. 29 below.

[28] The object of this procedure is to generate series of seasonally adjusted end-of-quarter stocks without any gaps or inconsistencies due to splicing of data series. The Federal Reserve System does not construct such series.

[29] In principle the data should reflect market valuation adjustments for all assets. The flow-of-funds series used here incorporate such adjustments only for equities and for corporate bonds issued by foreigners.

CB = holdings of long-term corporate bonds;
CM = holdings of commercial mortgages;
EQ = holdings of equities;
LA = holdings of liquid assets;
MB = holdings of municipal bonds;
RM = holdings of residential mortgages;
US = holdings of federal government securities;
r_{CB} = corporate bond yield (Moody's Aa new utility issues);
r_{CM} = commercial mortgage yield (ALIA series);
r_{CP} = commercial paper yield (prime 4–6-month paper);
r_{EQ} = equity yield (Standard and Poor's earnings/price ratio);
r_{MB} = municipal bond yield (Moody's Aaa);
r_{RM} = residential mortgage yield (new FHA insured loans);
r_{TB} = Treasury bill yield (3-month bills);
X_1 = lagged one-period percentage change of common stock prices (Standard and Poor's index);
X_2 = four-quarter moving average of X_1;
X_3 = lagged eight-quarter moving average percentage change of the consumer price index;
X_4 = lagged one-period percentage change of r_{CB}.

1. Life insurance companies:[31]

$$\begin{aligned}\Delta CB_t^L = {} & \underset{(-3.9)}{-3354} - \underset{(-1.7)}{1.177}\, \Delta W_t^L + \underset{(3.5)}{0.1465}\, W_{t-1}^L + \underset{(3.6)}{0.4550}\, (r_{CB,t} \cdot \Delta W_t^L)^* \\ & - \underset{(-2.5)}{0.003119}\, (r_{CB,t} \cdot W_{t-1}^L)^* - \underset{(-3.4)}{0.2624}\, r_{RM,t} \cdot \Delta W_t^L \\ & - \underset{(-2.5)}{0.008453}\, X_{1,t} \cdot \Delta W_t^L - \underset{(-1.9)}{0.1491}\, CB_{t-1}^L - \underset{(-3.7)}{0.1638}\, CM_{t-1}^L .\end{aligned}$$

$\bar{R}^2 = .80$ SE = 213 D-W = 1.56

---

[30] Because of the nonlinear way in which yield variables enter the model, as illustrated in equation (7), all such terms are products. The correct instrument to use in each such case, for purposes of deriving consistent estimators, is the first-stage fitted value of the entire product; this procedure is used here.

[31] The life insurance company equation also includes two dummy variables. The first dummy has unit value in 1969:IV and zero value in all other quarters; the flow-of-funds data show negative net purchases of corporate bonds by life insurance companies in this quarter (the only negative value for this series since these data began in 1953). The second dummy has unit value in 1973:I and zero value in all other quarters; the flow-of-funds data show an isolated, sharp one-quarter drop in net purchases of corporate bonds in this quarter.

2. Other insurance companies:[32]

$$\begin{aligned}\Delta CB_t^O = {} & \underset{(12.0)}{1696}\ WTR_t^O + \underset{(2.5)}{0.1401}\ (r_{CB,t}\cdot \Delta W_t^O)^* - \underset{(-1.8)}{0.1318}\ r_{MB,t}\cdot \Delta W_t^O \\ & - \underset{(-7.2)}{0.02259}\ X_{2,t}\cdot \Delta W_t^O - \underset{(-5.3)}{0.1390}\ CB_{t-1}^O - \underset{(-4.9)}{0.06440}\ US_{t-1}^O \\ & - \underset{(-15.8)}{0.1590}\ MB_{t-1}^O.\end{aligned}$$

$\bar{R}^2 = .92$ SE $= 66$ D-W $= 1.41$

3. Private pension funds:[33]

$$\begin{aligned}\Delta CB_t^P = {} & \underset{(3.3)}{0.2407}\ W_{t-1}^P + \underset{(6.5)}{0.1546}\ (r_{CB,t}\cdot \Delta W_t^P)^* - \underset{(-4.6)}{0.7157}\ X_{3,t}\cdot \Delta W_t^P \\ & - \underset{(-3.5)}{0.3730}\ CB_{t-1}^P + \underset{(3.6)}{0.2146}\ US_{t-1}^P - \underset{(-3.4)}{0.2493}\ EQ_{t-1}^P.\end{aligned}$$

$\bar{R}^2 = .67$ SE $= 198$ D-W $= 2.32$

4. State and local government retirement funds:

$$\begin{aligned}\Delta CB_t^S = {} & \underset{(-2.6)}{-1674} + \underset{(2.3)}{0.1144}\ W_{t-1}^S + \underset{(4.7)}{0.09882}\ (r_{CB,t}\cdot \Delta W_t^S)^* \\ & - \underset{(-3.9)}{0.4678}\ X_{3,t}\cdot \Delta W_t^S - \underset{(-3.9)}{0.01467}\ X_{4,t}\cdot \Delta W_t^S - \underset{(-2.0)}{0.1450}\ CB_{t-1}^S \\ & + \underset{(2.9)}{0.1098}\ US_{t-1}^S.\end{aligned}$$

$\bar{R}^2 = .83$ SE $= 156$ D-W $= 1.85$

5. Mutual savings banks:

$$\begin{aligned}\Delta CB_t^M = {} & \underset{(3.6)}{3819} + \underset{(6.5)}{0.1411}\ (r_{CB,t}\cdot \Delta W_t^M)^* - \underset{(-3.2)}{0.09626}\ r_{TB,t}\cdot \Delta W_t^M \\ & + \underset{(2.2)}{0.001588}\ r_{TB,t}\cdot W_{t-1}^M - \underset{(-2.7)}{0.002791}\ r_{CM,t}\cdot W_{t-1}^M \\ & - \underset{(-3.6)}{0.1219}\ CB_{t-1}^M - \underset{(-3.6)}{0.3465}\ US_{t-1}^M - \underset{(-2.9)}{0.1690}\ RM_{t-1}^M \\ & + \underset{(2.7)}{0.3024}\ CM_{t-1}^M.\end{aligned}$$

$\bar{R}^2 = .89$ SE $= 134$ D-W $= 2.12$

---

[32] The variable WTR^O is the trend component of the usual wealth variable W^O. The other insurance company equation also includes a dummy variable with unit value in 1973:II and zero value in all other quarters; the flow-of-funds data show an isolated

6. Households:

$$
\begin{aligned}
\Delta CB_t^H = {} & \underset{(4.4)}{0.03139}\,(r_{CB,t}\cdot \Delta W_t^H)^* + \underset{(3.5)}{0.0003370}\,(r_{CB,t}\cdot W_{t-1}^H)^* \\
& - \underset{(-1.6)}{0.02203}\, r_{EQ,t}\cdot \Delta W_t^H - \underset{(-3.5)}{0.01628}\, r_{CP,t}\cdot \Delta W_t^H - \underset{(-7.1)}{0.1167}\, CB_{t-1}^H \\
& + \underset{(2.2)}{0.008501}\, LA_{t-1}^H - \underset{(-3.9)}{0.006204}\, EQ_{t-1}^H .
\end{aligned}
$$

$\bar{R}^2 = .79$ SE = 496 D-W = 2.34

In all six equations the results provide support for the optimal marginal adjustment model of portfolio behavior developed in Sections I and III. First, in all six equations the combined own yield and flow term $(r_{CB,t}\cdot \Delta W_t)$ has a coefficient β_{ii} which is significantly different from zero, with the expected positive sign, at the 1 percent confidence level. The operational significance of this multiplicative term, which emerges from the combination of portfolio selection model (1) and portfolio adjustment model (6), is that the responsiveness of investors' bond purchases ΔCB_t to movements in the bond yield $r_{CB,t}$—that is, the partial derivative of ΔCB_t with respect to $r_{CB,t}$—depends positively on the current-quarter flow variable ΔW_t. For non–life insurance companies, private pension funds, state and local government retirement funds, and mutual savings banks, the $(r_{CB,t}\cdot \Delta W_t)$ term is the only appearance of the current value of the own yield in the bond demand equation. For life insurance companies and households, the term $(r_{CB,t}\cdot W_{t-1})$ also appears in the demand equation, with negative sign and positive sign, respectively.

For all five investor categories other than households, one aspect of these results may at first seem puzzling. Does the absence of the $(r_{CB,t}\cdot W_{t-1})$ term in the equations for four investor categories indicate that these investors' bond purchases are totally unresponsive to the own yield if the wealth flow is zero? Does the negative coefficient of the $(r_{CB,t}\cdot W_{t-1}^L)$ term in the life insurance company equation indicate that these companies' bond purchases respond negatively to the own yield if the wealth flow is less than a specific fraction of the existing stock of wealth? The answer to both questions is probably that it is not reasonable to extrapolate the results of estimating a model too far beyond the ob-

---

one-quarter shift to positive net purchases of corporate bonds by other insurance companies in this quarter, just in the midst of a sustained period of negative net purchases.

[33] The private pension fund equation also includes a dummy variable with unit value in 1971:IV and zero value in all other quarters; the flow-of-funds data show an isolated one-quarter shift to positive net purchases of corporate bonds by private pension funds in this quarter, just in the midst of a sustained period of negative purchases.

served range of the data variables. For life insurance companies, for example, the estimated partial derivative of ΔCB_t^L with respect to $r_{CB,t}$—that is, the value $(0.4550\,\Delta W_t^L - 0.003119\,W_{t-1}^L)$—is positive, given the actual values of ΔW_t^L and W_{t-1}^L, for all 56 quarters of the sample period.

All six estimated equations also conform to the portfolio behavior model developed in Sections I and III in that the lagged own-stock term CB_{t-1} in each case has a coefficient $-\theta_{ii}$ which is significantly different from zero, with the expected negative sign, at the 5 percent confidence level. In addition, the own-stock adjustment coefficients are probably of reasonable magnitude for a quarterly model, indicating that private pension funds (which typically manage their portfolios actively and also invest primarily in publicly traded securities) adjust their bond holdings significantly more rapidly than either households or the other four categories of institutional investors.

The remaining aspects of the empirical results require little comment here. The $496 million standard error for households is by far the largest among the six estimated demand equations, but this result is not surprising in light of the high variance of the data for this sector which is a residual element of the flow-of-funds accounts. The few nonyield variables, X_h, $h = 1, \ldots, 4$, represent a minimal allowance for some of the more important expectational variables which influence investors' behavior in the corporate bond market; more thorough investigation of these influences remains as an object of further research.

81606
Date 26.5.81

V. Simulation Results

As the discussion of Section IV indicates, the empirical results for the six bond demand equations provide support for the hypotheses embodied in the underlying optimal marginal adjustment model of portfolio behavior, including the role of financial flow variables in influencing short-term portfolio adjustments. It remains to examine the performance of the structural model as a whole in explaining short-run movements in the model's jointly determined variable—the long-term bond yield itself.

Figure 3 and Part A of table 3 present simulation results for the seven-equation structural model consisting of the six demand-for-bonds equations presented in Section IV and the market-clearing identity

$$\Delta CB_t^L + \Delta CB_t^O + \Delta CB_t^P + \Delta CB_t^S + \Delta CB_t^M + \Delta CB_t^H + \Delta CB_t^{EXD} = \Delta CB_t^{SUP}, \tag{8}$$

where ΔCB_t^{EXD} is the net purchases of corporate bonds by investors taken to be exogenous, and ΔCB^{SUP} is the (exogenous) total net supply of bonds by all issuers. The simulation period is 1960:I–1973:IV. This simulation is fully dynamic in that, after the first quarter of the simulation period,

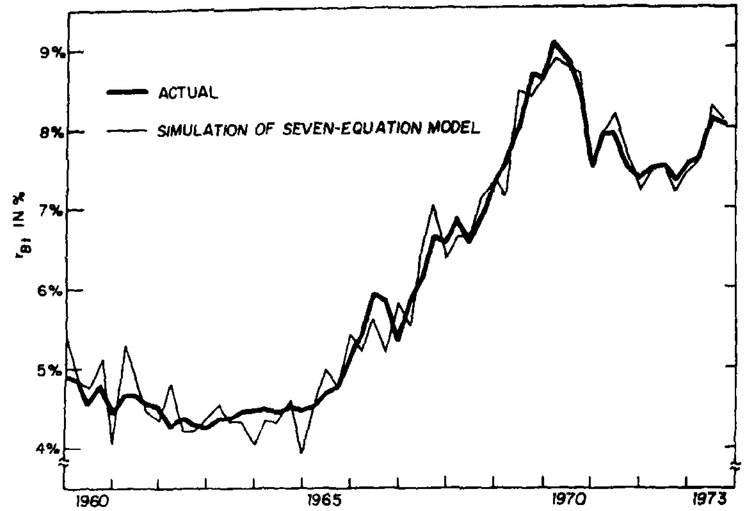


FIG. 3.—Dynamic simulation results for Aa utility new issue yield

the solution uses internally generated values for the lagged own-stock variables CB_{t-1} in all six equations (as well as internally generated values of $r_{CB,t-1}$ and $r_{CB,t-2}$ in variable $X_{4,t}$).[34]

Part A of table 3 shows the mean and root-mean-square simulation errors for the Aa new-issue utility bond yield r_{CB} and the six net bond purchases variables ΔCB. Figure 3 plots the actual values of r_{CB} against the simulated values. These simulation results show no significant bias for any of the seven endogenous variables. For the six bond purchases variables, the root-mean-square errors are about in line with the standard errors of the estimated equations, indicating that the errors made by individual equations have no observable tendency to compound one another in a dynamic context. For the bond yield itself, the 0.27 percent root-mean-square error is about comparable to the fit achieved by other researchers using the reduced-form term-structure approach.[35] This

---

[34] To provide at least some limited precaution against the multiple solution problem in simulations of systems of nonlinear simultaneous equations (see Friedman [1971]), the simulations reported here were performed several times with various orderings of the individual equations; the results were unchanged.

[35] Modigliani and Shiller's (1973) preferred equation had SE = .13 percent for the less volatile Aaa yield over the sample period 1955:III–1971:II. Reestimating the Modigliani-Shiller equation using the Aa yield and the 1960:I–1973:IV sample period leads to an equation with SE = .22 percent but with the coefficients of the distributed lag on the short-term yield not significantly different from zero. Feldstein and Eckstein's (1970) preferred equation had SE = .09 percent for the Aaa yield over the sample period 1954:I–1969:II. Reestimating the Feldstein-Eckstein equation using the Aa yield and the 1960:I–1973:IV sample period leads to an equation with SE = .29 percent.

TABLE 3

DYNAMIC SIMULATION RESULTS FOR 1960:I–1973:IV

A. SEVEN-EQUATION STRUCTURAL MODEL WITH BOND SUPPLY EXOGENOUS

Variable	Mean Error	Root-Mean-Square Error
r_{CB}	0.01	0.27
ΔCB^L	−1.4	165
ΔCB^O	0.1	60
ΔCB^P	1.3	202
ΔCB^S	0.1	192
ΔCB^M	−0.6	125
ΔCB^H	0.4	332

B. NINE-EQUATION STRUCTURAL MODEL WITH BOND SUPPLY ENDOGENOUS

Variable	Mean Error	Root-Mean-Square Error
r_{CB}	−0.01	0.21
ΔCB^L	1.0	181
ΔCB^O	0.4	64
ΔCB^P	1.0	213
ΔCB^S	1.9	156
ΔCB^M	0.9	127
ΔCB^H	1.8	423
ΔCB^N	13.2	405
ΔCB^F	−6.2	197

NOTE.—r_{CB} in percent, all ΔCB in $ million.

within-sample performance seems quite creditable, especially since the methodology of the structural model does not estimate an unrestricted equation directly for the bond yield but, instead, implies an equation for the bond yield which is restricted by the underlying structural hypotheses about portfolio behavior.[36] In addition, since the focus of this paper is the role of financial flow variables in portfolio behavior, the six estimated demand equations exclude elaborate distributed-lag representations of holding-period yield expectations as well as some nonyield explanatory variables suggested in Friedman (1974).

As is clear from close inspection of figure 3, the simulated values of the bond yield track the actual values more closely in the second half of the period (RMSE = 0.23 percent for 1967:I–1973:IV) than during the first

---

Feldstein and Chamberlain's (1973) preferred equation had SE = .21 percent for the Aaa yield over the sample period 1954:I–1971:I.

[36] This point is especially relevant to the presence of other long-term yields in the estimated bond demand equations for several investor groups. Including other long-term yields as independent variables in an unrestricted equation with the bond yield as dependent variable would presumably increase greatly such an equation's fit but, in so doing, would in part incorporate spurious correlations. In the context of the structural model, however, the contribution of other long-term yields is restricted to their role in influencing the net purchases variables.

half (RMSE = 0.31 percent for 1960:I–1966:IV); the appropriate test based on the F-statistic ($F = 1.71$) indicates that this difference is significant at the 10 percent confidence level but not at the 5 percent confidence level. Given the "inverted" nature of the system of equations solving for the bond yield, the difference may be due to the greater volatility of the bond yield in more recent years. If so, the ability to track well during periods of greater yield volatility is an attractive feature of the structural model.

Because it takes the total bond supply to be exogenous, this simulation of the seven-equation model cannot capture the supply-demand interaction of a full structural model of the market determination of long-term interest rates. Part B of table 3 presents results for an analogous simulation of the nine-equation structural model consisting of the six demand-for-bonds equations presented in Section IV,[37] the two supply-of-bonds equations presented in Friedman (1976), and the market-clearing identity

$$\Delta CB_t^L + \Delta CB_t^O + \Delta CB_t^P + \Delta CB_t^S + \Delta CB_t^M + \Delta CB_t^H + \Delta CB_t^{EXD} = \Delta CB_t^N + \Delta CB_t^F + \Delta CB_t^{EXS}, \tag{9}$$

where the right-hand-side variables are the net supplies of bonds by (in order) nonfinancial business corporations, finance companies, and issuers taken to be exogenous.[38]

As the r_{CB} error mean and root-mean-square in Part B of table 3 show, the nine-equation structural model with endogenous bond supplies outperforms the seven-equation model in its ability to track the historical movements of the own yield on long-term bonds. There is no significant bias, and the 0.21 percent root-mean-square error seems most creditable.[39] For the eight bond purchases and issues variables, a small (but not significant) upward bias in net issues by nonfinancial corporations leads to a small (again not significant) downward bias in net issues by finance companies and even smaller uniformly positive upward biases in net purchases by all six investor groups.

Either by themselves or in conjunction with equations for bond supply, therefore, the six demand-for-bonds equations presented in Section IV form the basis of a structural model of long-term interest rate determination which performs well in dynamic simulation tests.

---

[37] The six demand-for-bonds equations actually used in this simulation were reestimated to allow for the change in the model's set of exogenous variables. The reestimated coefficients differed very little from those reported in Section IV.

[38] As of the end of 1974, nonfinancial business corporations and finance companies, the two endogenous categories of bond issuers, together accounted for over 91 percent of all outstanding corporate bonds issued in the United States.

[39] See again the comparisons in n. 35 above.

VI. Conclusions

Two principal conclusions emerge from the empirical work presented in this paper.

First, the individual-equation estimation results support the "optimal marginal adjustment" model of investors' portfolio behavior in the presence of transactions costs which are greater for the reallocation of existing asset holdings than for the initial allocation of new wealth increments. In particular, they support the hypothesis, which underlies the optimal marginal adjustment model, that investable cash flow variables are a significant determinant of investors' short-run asset demands and hence of asset yields as well.

Second, on a methodological level, the full-model dynamic simulation results support the use of a structural model of the determination of long-term interest rates. The implied expression for the long-term bond yield, which is restricted by the estimated structural demand and supply equations, fits the data about as well as do previously developed unrestricted reduced-form term-structure equations. The six investor categories' demand-for-bonds equations perform about as well in dynamic simulations as in individual-equation tests. Given the advantage of a framework consisting of structurally specified demand and supply equations for testing explicit hypotheses about portfolio behavior, therefore, further research on structural models of long-term interest rate determination appears warranted on the basis of these full-model results.

References

Anderson, W. H. Locke. *Corporate Finance and Fixed Investment.* Boston: Graduate School Bus. Admin., Harvard Univ., 1964.

Ando, Albert, and Klein, Lawrence R. "Progress of Specifications and Estimation of a Model for Flow of Funds Accounts." Mimeographed. Univ. Pennsylvania, 1975.

Ando, Albert, and Modigliani, Franco. "Some Reflections on Describing Structures of Financial Sectors." In *The Brookings Model: Perspective and Recent Developments,* edited by G. Fromm and L. R. Klein. Amsterdam: North-Holland, 1975.

Board of Governors of the Federal Reserve System. *Flow of Funds Accounts 1945–72.* Washington: Board Governors Federal Reserve System, 1973.

Bosworth, Barry, and Duesenberry, James S. "A Flow of Funds Model and Its Implications." In *Issues in Federal Debt Management.* Boston: Federal Reserve Bank of Boston, 1973.

Brainard, William C., and Tobin, James. "Pitfalls in Financial Model-Building." *A.E.R.* 57 (May 1968): 99–122.

Brundy, James M., and Jorgenson, Dale W. "Efficient Estimation of Simultaneous Equations by Instrumental Variables." *Rev. Econ. and Statis.* 53 (August 1971): 207–24.

Culbertson, J. M. "The Term-Structure of Interest Rates." *Q.J.E.* 71 (November 1957): 485–517.

De Leeuw, Frank. "A Model of Financial Behavior." In *The Brookings Quarterly Econometric Model of the United States*, edited by J. S. Duesenberry et al. Chicago: Rand McNally, 1965.

Duesenberry, James S. "The Portfolio Approach to the Demand for Money and Other Assets." *Rev. Econ. and Statis.* 65 (February 1963): 9–24.

Fair, Ray C., and Malkiel, Burton G. "The Determination of Yield Differentials between Debt Instruments of the Same Maturity." *J. Money, Credit and Banking* 3 (November 1971): 733–49.

Feldstein, Martin, and Chamberlain, Gary. "Multimarket Expectations and the Rate of Interest." *J. Money, Credit and Banking* 5 (November 1973): 873–902.

Feldstein, Martin, and Eckstein, Otto. "The Fundamental Determinants of the Interest Rate." *Rev. Econ. and Statis.* 52 (November 1970): 363–75.

Friedman, Benjamin M. "Econometric Simulation Difficulties: An Illustration." *Rev. Econ. and Statis.* 53 (November 1971): 381–84.

———. "A Structural Model of the Long-Term Corporate Debt Market." Mimeographed. Harvard Univ., 1974.

———. "Substitution and Expectation Effects on Long-Term Borrowing Behavior and Long-Term Interest Rates." Mimeographed. Harvard Univ., 1976.

Friedman, Milton. "The Quantity Theory of Money: A Restatement." In *Studies in the Quantity Theory of Money*, edited by M. Friedman. Chicago: Univ. Chicago Press, 1956.

Fromm, Gary, and Sinai, Allen. "A Policy Simulation Model of Deposit Flows, Mortgage Sector Activity, and Housing." Mimeographed. Data Resources Inc., 1975.

Goldfeld, Stephen M. *Commercial Bank Behavior and Economic Activity.* Amsterdam: North-Holland, 1966.

———. "An Extension of the Monetary Sector." In *The Brookings Model: Some Further Results*, edited by J. S. Duesenberry et al. Chicago: Rand McNally, 1969.

Hendershott, Patric H. "A Flow-of-Funds Model: Estimates for the Non-Bank Finance Sector." *J. Money, Credit and Banking* 3 (November 1971): 815–32.

Hendershott, Patric H., and Lemmon, Richard C. "A Flow of Funds Model: First Estimates and Forecasts." *American Statistical Association Journal Proceedings* (1973), business and economics section, pp. 112–21.

Hicks, John R. *Value and Capital.* London: Oxford Univ. Press, 1939.

Kessel, Reuben A. "The Cyclical Behavior of the Term Structure of Interest Rates." In *Essays on Interest Rates*, edited by J. Guttentag. Vol. 2. New York: Columbia Univ. Press, 1971.

Kopcke, Richard W. "An Econometric Analysis of Household Demand Behavior for Financial Assets." Ph.D. dissertation, Harvard Univ., 1973.

Ladenson, Mark L. "Pitfalls in Financial Model Building: Some Extensions." *A.E.R.* 61 (March 1971): 179–86.

Malkiel, Burton Gordon. *The Term Structure of Interest Rates.* Princeton, N.J.: Princeton Univ. Press, 1966.

Meiselman, David. *The Term Structure of Interest Rates.* Englewood Cliffs, N.J.: Prentice-Hall, 1962.

Modigliani, Franco. "The Dynamics of Portfolio Adjustment and the Flow of Savings through Financial Intermediaries." In *Savings Deposits, Mortgages and Housing*, edited by E. M. Gramlich and D. M. Jaffee. Lexington, Mass.: Heath, 1972.

Modigliani, Franco, and Shiller, Robert J. "Inflation, Rational Expectations, and the Term Structure of Interest Rates." *Economica* 40 (February 1973): 12–43.

———. "New Issues and the Term-Coupon Structure of Bond Yields." Mimeographed. M.I.T., 1974.

Modigliani, Franco, and Sutch, Richard. "Innovations in Interest Rate Policy." *A.E.R.* 56 (May 1966): 178–97.

———. "Debt Management and the Term Structure of Interest Rates: An Empirical Analysis of Recent Experience." *J.P.E.* 75, no. 4 (August 1967): 569–89.

Shapiro, Eli, and Wolf, Charles R. *The Role of Private Placements in Corporate Finance.* Boston: Harvard Univ., Graduate School Bus. Admin., 1972.

Silber, William L. *Portfolio Behavior of Financial Institutions.* New York: Holt, Rinehart & Winston, 1970.

Smith, Gary. "Pitfalls in Financial Model Building: A Clarification." *A.E.R.* 65 (June 1975): 510–16.

Smith, Gary, and Brainard, William. "Estimation of the Savings Sector in a Disequilibrium Model." Mimeographed. Yale Univ., 1974.

The Demand for Quality-adjusted Cash Balances: Price Uncertainty in the U.S. Demand for Money Function

Benjamin Klein

University of Chicago Law School and University of California, Los Angeles

This paper investigates the importance of uncertainty regarding the rate of price change as an argument in the long-run money demand function. Increased inflation uncertainty is assumed to lower the stream of monetary services yielded by a given level of real cash balances. The effects on money demand of such changes in the "quality" of money are, in general, theoretically indeterminate. If it is assumed, however, that the monetary service flow is proportional to the real money stock and that the demand for money is interest inelastic, then the predicted relationship between price uncertainty and money demand is unambiguously positive. The empirical findings of this paper, where price uncertainty is operationally measured by the variability of the rate of price change, strongly confirm this positive relationship. These results have important implications for the theory of inflation, the optimum quantity of money, and the potential government tax revenue from money creation.

Most empirical work on the demand for money function has either implicitly or explicitly considered money to be a durable good yielding an unspecified flow of utility-providing services. The demand for money is thus a stock demand derived from a flow demand, and its treatment in theoretical and empirical studies should be no different from the treatment given to the demand for other producer and consumer durables. It is puzzling, then, that while empirical studies of the demand for other

I am indebted to Jack Carr, Michael Darby, Jacob Frenkel, Milton Friedman, Thomas Heagy, Rob Kolson, Roger Kormendi, Laura LaHaye, Sean Mooney, Anna Schwartz, and participants at money workshops at UCLA and the University of Chicago for useful comments. Extremely able research assistance was provided by Laura LaHaye. An earlier draft of this paper was written while I was a Research Fellow at the National Bureau of Economic Research during 1971–72. I am also grateful to the Foundation for Research in Economics and Education and to the program in Law and Economics at the University of Chicago Law School for research support.

[*Journal of Political Economy*, 1977, vol. 85, no. 4]

© 1977 by The University of Chicago. All rights reserved.

durable goods have explicitly defined the quantity of the stock in terms of the service flow (see, for example, Harberger 1960) no attempt has yet been made to standardize the money stock in terms of its monetary service flow other than to deflate the nominal stock by the price level. Clearly, there are other factors which affect this stock-flow relationship. It is the purpose of this paper to investigate the empirical importance of changes in the "quality" of cash balances where quality is defined as the flow of monetary services yielded per real dollar. In general, time-series studies of the demand for money have implicitly assumed this quality to remain unchanged so that, for example, a real dollar in 1870 is assumed to be equivalent in terms of monetary services to a real dollar in 1970.

In this paper, the quality of a given level of real money holdings is simply assumed to be positively related to the level of price predictability. To explain precisely why this is true requires a consideration of the exact role of money in an exchange economy; an extended discussion of this important issue, however, is beyond the scope of this paper. In Section I, the theory is presented, and the demand for money is considered similarly to the demand for any productive input. In Section II, quality is operationally defined in terms of price variability. This variable is shown to enter the U.S. demand for money function positively over the last century and to provide a possible explanation for the sharp 1967–71 break in the postwar growth of velocity. These results, however, are suspect because of the presence of first-order serial correlation in the residuals. In Section III, therefore, a time-series analysis of the relationship is presented. The results indicate that, after correcting for the serial correlation, price variability enters the demand for money function positively and significantly. Finally, Section IV presents implications of these results for monetary theory.

I. Theory

Let N be the flow of monetary services an individual derives per unit time from his holdings of real cash balances, and let the relationship between the stock and the flow be represented by

$$N = N(M/P, \beta) \tag{1}$$

where M/P is his stock of real cash balances and β is a shift variable representing its quality.[1] Equation (1) can be viewed as the representative individual's production function for monetary services with M/P and β assumed to have declining marginal productivities.[2]

---

[1] To avoid misunderstanding I should explicitly note that the meaning of "quality" in this paper is different from its meaning in my article, "The Competitive Supply of Money" (1974*b*). There quality was defined to be negatively related to the (mean) anticipated inflation rate in terms of the money. This should be thought of as the anticipated quality of a *nominal* unit of money. On the other hand, β refers to the quality of a

In a recent article (1974*a*) I define "the price of money," P_M (or the rental price of the monetary service stream from a dollar of money), by $i - r_M$, the marginal pecuniary alternative cost per unit time of holding a dollar of money where i is the return on an asset yielding no monetary service and r_M is the pecuniary return on money. In equilibrium, P_M will equal the value of the monetary services from a marginal dollar of money, or

$$P_M = (i - r_M) = N_{(M)} \cdot P_N \tag{2}$$

where $N_{(M)}$ is the marginal productivity of money in terms of monetary services and P_N is the price of monetary services. From my simple model, the individual's desired monetary service stream can be derived as a function of y_p, the real permanent income or wealth constraint of the individual and (P_N/P), the price of a unit rate of monetary services relative to commodity services. Thus

$$N^d = f[y_p, (P_N/P)] \tag{3}$$

where (P_N/P), the real or relative price of monetary services, equals $(i - r_M)/N_{(M/P)}$,[3] and where $\partial N^d/\partial y_p > 0$ and $\partial N^d/\partial(P_N/P) < 0$. Given this demand function, the individual's production function for monetary services then determines his derived demand for real cash balances.

Current empirical work on the demand for money implicitly assumes that the monetary service flow is proportional to the real stock of cash balances held and that the proportionality constant does not change over time.[4] In that case, the stock of real cash balances is a perfect proxy for the flow of monetary services, and $(i - r_M)$ can be used as a measure of the relative price of monetary services. In other words, the demand for

---

real unit of money, or what I called in my earlier article the "confidence" in the future exchange value of the money, assumed there to be negatively related to the variance of the anticipated rate of change in prices. Other variables that could influence the monetary service flow for a given level of real cash balances include, for example, the number of banking offices per capita and the composition of currency denominations.

[2] It is assumed here for expositional simplicity that there are no alternative sources of monetary services, so that the effects of shifts in β on the demand for real cash balances can be clearly isolated. The empirical work in this study, however, considers the return on a monetary substitute asset as a relevant determinant of money demand. If such a substitute asset is introduced into the production function, the theoretical results derived below are unchanged, that is, $\partial(M/P)/\partial\beta$ remains ambiguous. These results can be obtained from me upon request.

[3] Since each individual demander is assumed to be a perfect competitor who takes the price level as given, i.e., $\partial P/\partial M = 0$, then $\partial(M/P)/\partial M = 1/P$ and therefore $N_{(M)} = N_{(M/P)} \cdot (1/P)$.

[4] For example, Friedman (1969) "interpret[s] money as a durable good held for the services it renders and [which yields] a flow of services proportional to the stock . . ." (p. 119).

money can be written as

$$(M/P)^d = g[y_p, (i - r_M)]. \tag{4}$$

This specification may be appropriate for very short-run problems, but when estimating a long-run demand for money relationship movements along or shifts of the production function for monetary services may be important; hence, the derived demand for real cash balances will be more generally determined by the demand for monetary services, equation (3), and the production function for monetary services, equation (1). Given, from (1), that $N/(M/P)$ equals the average output of monetary services per dollar of real balances held, then $N^d = N/(M/P) \cdot (M/P)^d$, or

$$(M/P)^d = h[y_p, (P_N/P), N/(M/P)] \tag{5}$$

where

$$\partial(M/P)^d/\partial y_p > 0, \qquad \partial(M/P)^d/\partial(P_N/P) < 0$$

and

$$\partial(M/P)^d/\partial[N/(M/P)] < 0.$$

The last partial indicates that for a given y_p and (P_N/P) (that is, for a given N^d), a higher average "productivity" of real balances implies a lower demand for money. Fewer real cash balances are now necessary to produce the desired monetary service flow.

A shift in the production function for monetary services therefore has ambiguous effects on the demand for real balances. An increase in β, representing an upward shift in this production function, implies an increase in $N/(M/P)$ and therefore a decrease in the demand for money. But the shift may also imply an increase in $\partial N/\partial(M/P)$ and therefore a decrease in (P_N/P) and an increase in the demand for money (if $[M/P]$ and β are complementary factors throughout). In other words, an increase in the quality of real money implies (1) that the average productivity of real money rises and less real money is now necessary to produce a given monetary service flow and (2) that the marginal productivity of real money may also rise, decreasing the price of monetary services and hence increasing the service flow demanded.

The two distinct effects of an increase in β on $(M/P)^d$ are, ignoring any real income effects, (1) a decrease in $(M/P)^d$ because (M/P) and β (quantity and quality) are substitutes in the production of N (that is, along an N isoquant $d[M/P]/d\beta < 0$); (2) a possible increase in $(M/P)^d$ because (M/P) and N are complements in consumption (that is, $\partial[M/P]^d/\partial P_N < 0$, and $\partial P_N/\partial\beta$ is likely to be < 0). Perhaps a more intuitive notion of these two effects is obtained if β is considered, for example, as a measure of the efficiency of the transactions process. This variable, which Irving Fisher earlier emphasized as an important determinant of the demand for money, has now become especially significant

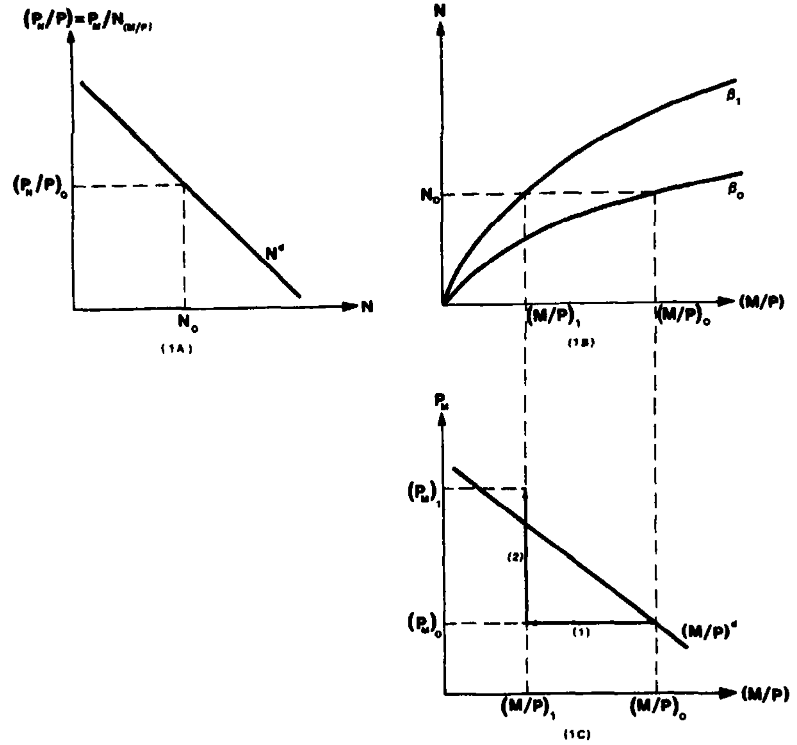


FIG. 1

in light of recent discussion concerning the implications for the demand for money of the growth of credit-card usage and other movements toward a "checkless" society. These technological innovations permit one to get more monetary services out of the same amount of cash, so they produce a decrease in the demand for money (effect 1), but they also, in general, make monetary services cheaper, which increases the demand for money (effect 2). The net effect, then, is ambiguous, and it is thus presumptuous to conclude, as did Fisher and as is often done now, that improvements in transportation and communication over time will necessarily lead to a secular rise in the velocity of money.

A change in β is depicted graphically (fig. 1) by a rotation in the demand for cash balances schedule rather than an unambiguous shift in one direction or the other. Given the demand for monetary services schedule (eq. [3]) in figure 1*A* and the monetary service production schedule (eq. [1]) in figure 1*B*, the demand for money schedule is traced out in figure 1*C*. For example, given a particular real cost of monetary services, $(P_N/P)_0$, 1A indicates the desired monetary service flow, N_0; 1B then indicates the level of cash balances necessary to produce that service flow, $(M/P)_0$, and its implicit marginal productivity, $N_{(M/P)_0}$, which, when multiplied by $(P_N/P)_0$ yields the real cost of holding the money $(P_M)_0$—or one point on the demand for money schedule in 1C, $[(M/P)_0, (P_M)_0]$, corresponding to the original point on the demand for money services schedule. Assume now that β rises from β_0 to β_1. At the same $(P_N/P)_0$, N^d is unchanged at N_0, but the corresponding point in 1C will shift. First, at the old implicit cost of holding money fewer real cash balances, represented by $(M/P)_1$, will now be necessary to produce the given N_0 (this is the average productivity effect 1). Second, since, from 1B, $N_{(M/P)}$ at N_0 is now higher, the implicit P_M at $(P_N/P)_0$ is also now higher than $(P_M)_0$, say $(P_M)_1$ (this is the marginal productivity effect 2). The corresponding point on the new demand schedule, $[(M/P)_1, (P_M)_1]$, therefore, is to the northwest of the old point. Since at $(P_N/P) = 0$ only the first horizontal leftward shift (the substitution-in-production effect) is present, while at $N^d = 0$ only the second vertical upward shift (the complementarity-in-consumption effect) is present, only if the demand for monetary services is unbounded can the sign of $\partial(M/P)/\partial\beta$ possibly be unambiguous.[5] More generally, the demand for money schedule will rotate clockwise with an increase in β.

The particular direction of effect on $(M/P)^d$ of a change in β can be determined by considering the change in the real value of the marginal product of money, $VMP_{(M/P)}$, as β changes, where

$$VMP_{(M/P)} \equiv N_{(M/P)} \cdot [P_N/P] \tag{6}$$

and where $VMP_{(M/P)}$, by the law of marginal productivity, is equal in equilibrium to $(i - r_M)$. The change in the demand price at a given output from a change in quality is therefore given by

$$\left.\frac{\partial VMP_{(M/P)}}{\partial\beta}\right|_{(M/P)} = \underset{(+)}{\overset{(1)}{N_{(M/P)}}}\ \underset{(-)}{\overset{(2)}{\frac{\partial(P_N/P)}{\partial N}}}\ \underset{(+)}{\overset{(3)}{\frac{\partial N}{\partial\beta}}} + \underset{(+)}{\overset{(4)}{(P_N/P)}}\ \underset{(?)}{\overset{(5)}{\frac{\partial N_{(M/P)}}{\partial\beta}}}. \tag{7}$$

---

[5] A necessary condition for a bounded monetary service demand schedule is that $N^d < \infty$ at $(P_N/P) = 0$. The commonly used logarithmic functional form for the demand for money is unbounded at a zero rate of interest. But once risk considerations associated with the future real value of money are introduced, then even at zero (or negative) net cost of holding money an optimum portfolio will not consist entirely of money. As long as the net return on money has a correlation with the return on real assets which is less than unity, risk-averse individuals will diversify and hold a fraction of their wealth in common stocks and other real assets, even if the expected return is lower. The Keynesian analysis often cited to justify the use of a logarithmic demand for money function considers nominally denominated bonds as the sole alternative to money and therefore can properly ignore the risk associated with unanticipated price level changes. Riskiness refers only to the capital value changes associated with interest rate changes, and money is therefore correctly assumed in this narrow context to be a perfectly riskless asset (see also n. 11 below).

Under each of the five terms is the sign for the direction of change. Term 2 is the slope of the demand for monetary service function, term 3 is what I have referred to as the "average productivity" factor, and term 5 is the "marginal productivity" factor. Terms 1 and 4 are the "weights" placed on these two effects (for example, if $[P_N/P] = 0$ the marginal effect is zero, whereas if $N_{[M/P]} = 0$ the average effect is zero). It is clear that $1 \cdot 2 \cdot 3$, which I have called the substitution-in-production effect, is necessarily negative, but $4 \cdot 5$, which I have called the complementarity-in-consumption effect, is ambiguous in sign. If $\partial N_{(M/P)}/\partial\beta$ is greater than zero the consumption effect is positive, and it is possible that it could outweigh the production effect and $\partial VMP_{(M/P)}/\partial\beta$ could be greater than zero. The total net effect of a change in quality on the demand for money is therefore ambiguous.[6]

If, however, the monetary service flow is proportional to the stock of real cash balances,

$$N = \beta(M/P), \tag{1'}$$

then it can be easily shown that the point on the demand schedule around which the demand for money rotates is of unitary price elasticity, and $\partial(M/P)^d/\partial\beta$ will be greater than zero in the elastic portion of the demand schedule and less than zero in the inelastic portion of the demand schedule.[7] Since the demand for money is commonly estimated to be price

---

[6] A somewhat more appropriate way to formulate this is to consider the change in the quantity demanded at a given price, P_M, from a change in quality. That is, for a given $(i - r_M)$ (implying therefore in equilibrium no change in $VMP_{[M/P]}$), allow (M/P) to vary as β changes:

$$\begin{aligned}\frac{\partial VMP}{\partial\beta} = 0 = {} & N_{(M/P)}\frac{\partial(P_N/P)}{\partial N}\cdot\frac{\partial N}{\partial\beta} + (P_N/P)\frac{\partial N_{(M/P)}}{\partial\beta}\\ & + N_{(M/P)}\frac{\partial(P_N/P)}{\partial N}\,\frac{\partial N}{\partial(M/P)}\,\frac{\partial(M/P)}{\partial\beta} + (P_N/P)\frac{\partial N_{(M/P)}}{\partial(M/P)}\,\frac{\partial(M/P)}{\partial\beta},\end{aligned}$$

or

$$\frac{\partial(M/P)}{\partial\beta} = \frac{-\left[\overset{(+)}{N_{(M/P)}}\,\overset{(-)}{\frac{\partial(P_N/P)}{\partial N}}\cdot\overset{(+)}{\frac{\partial N}{\partial\beta}} + \overset{(+)}{(P_N/P)}\,\overset{(?)}{\frac{\partial N_{(M/P)}}{\partial\beta}}\right]}{\left[\underset{(+)}{N_{(M/P)}}\,\underset{(-)}{\frac{\partial(P_N/P)}{\partial N}}\,\underset{(+)}{\frac{\partial N}{\partial(M/P)}} + \underset{(+)}{(P_N/P)}\,\underset{(-)}{\frac{\partial N_{(M/P)}}{\partial(M/P)}}\right]}. \tag{7'}$$

Since the denominator is negative and the numerator is merely the negative of expression (7), the direction of change results for (7′) are identical with those given in (7).

[7] From production function (1′) and the demand for monetary services, $N^d = f[(P_N/P)]$, where $(P_N/P) = P_M/\beta$, we can therefore write the demand for money function as $(M/P)^d = (1/\beta)f[(P_M/\beta)]$, and $\partial(M/P)^d/\partial\beta = -(1/\beta^2)f[(P_M/\beta)] - (1/\beta)f'[(P_M/\beta)]\cdot(P_M/\beta^2)$. Substituting (P_N/P) for (P_M/β) and factoring we get $\partial(M/P)^d/\partial\beta = -(1/\beta^2)f[(P_N/P)](1-\eta_N)$, where $\eta_N = -(P_N/P)/f[(P_N/P)]\cdot f'[(P_N/P)]$. And therefore $\partial(M/P)^d/\partial\beta \gtreqless 0$ as $\eta_N \lesseqgtr 1$, or, since $N = \beta(M/P)$ and $\eta_N = \eta_{(M/P)}$ at every P_M, $\partial(M/P)^d/\partial\beta \gtreqless 0$ as $\eta_{(M/P)} \lesseqgtr 1$. For example, given (1′) and a rectangular hyperbola N^d function, the $(M/P)^d$ function is also rectangular hyperbola ([1′] merely changes the scale of our axes), and a change in β has no effect on $(M/P)^d$. (When β

inelastic,[8] an increase in quality can be expected to decrease the demand for money if production function (1′) holds.[9]

II. Preliminary Empirical Results

A. Model to Be Estimated

Since N is unobserved the demand for monetary services, equation (3), cannot be directly estimated. The preceding theoretical framework demonstrates, however, that a viable alternative is to estimate the stock demand for money, including as arguments those variables which can be expected to influence the relationship between the stock and the monetary service flow. The model we will empirically estimate is therefore

$$\log (M/P) = a_0 + a_1 \log y_p + a_2 r_S + a_3 r_L + a_4 r_M + a_5 \beta + u_t \qquad (8)$$

where (M/P) is real cash balances, y_p is real permanent income, r_S is a short-term (monetary substitute) interest rate, r_L is a long-term interest rate, r_M is the rate of return on money, and β represents the shift variables which influence the money stock–monetary service flow relationship.

The short-term perfectly illiquid rate, i, that is used as a theoretical benchmark in defining the monetary service flow from money is not included in this model because of the difficulty of obtaining a market

---

increases the leftward horizontal shift is equal to the upward vertical shift. Although the actual point on the $[M/P]^d$ schedule corresponding to a particular point on the N^d schedule shifts, there is no rotation of the entire schedule.) More generally, if the production function is of the general form (1), then $N^d = f[(P_N/P)]$ implies $(M/P)^d = [(M/P)/N] \cdot f[(P_M/N_{M/P})]$ and $\partial(M/P)^d/\partial\beta = [(M/P) f N_\beta / N^2] + [(M/P)/N] f' \times [P_M N_{M/P,\beta} / N^2_{(M/P)}] = [(M/P)/N] f[(N_\beta/N) + \eta_N (N_{M/P,\beta}/N_{M/P})]$. Therefore, $\partial(M/P)^d/\partial\beta$ will equal zero if $\eta_N = -(N_\beta \cdot N_{M/P}/N \cdot N_{M/P,\beta})$.

[8] If interest payments on money are highly correlated with short-term interest rates, the cross-price elasticity of demand for money is likely to be small. From my earlier study (1974*a*), the elasticity of demand for money with respect to $(r_S - r_M)$, evaluated at the sample means over the 1880–1970 period, was -0.347. This result and the results reported below are based upon a semilogarithmic functional form, which implies a variable interest elasticity demand function. Although most previous studies also find the demand for money to be interest inelastic, some impose a constant elasticity functional form on their estimates and all ignore the implicit pecuniary return on money.

[9] In addition, on purely inventory-theoretic–type grounds, the demand for money may be negatively related to quality (or positively related to price uncertainty). If, for example, the store of value precautionary motive has to do with maintaining a particular level of real consumption without any "forced" sale of illiquid assets, then a larger variance of prices would require larger real money balances to maintain any probability level of not having to make a forced sale. An analytically similar argument holds for the transactions motive. If an individual did not know whether the cup of coffee he planned to purchase for breakfast tomorrow morning would cost 10¢ or $10.00, his demand for real balances would certainly be larger. This, however, assumes that planned purchases are exogenous, while risk-aversion considerations would imply that individuals would increase current consumption and decrease money holdings to avoid next period's price uncertainty.

measure of this rate. The short-term rate on monetary substitutes is instead considered as the "alternative cost" of holding money, with $(r_S - r_M)$ now considered as the implicit "price" of money. Changes in r_S are likely to be highly correlated with changes in i over time, and therefore the failure to include i may not pose any serious problem empirically. The long-term interest rate is included as a measure of expectations regarding future short-term interest rates. If there are costs of moving in and out of cash and short-term assets, the expected behavior of future short-term rates will influence current money demand. Although the theory may therefore suggest a functional form which uses $(r_S - r_M)$ and r_L as arguments, empirically it often pays to use the extra 1 df to run the relationship in the simple unconstrained r_S, r_L, r_M interest rate functional form.[10] Consequently, the estimation procedures in this paper are based on this simpler unconstrained functional form.

B. An Operational Measure of Quality

The difficult empirical problem is to define operationally the variable β, and this requires some consideration of what "monetary services" actually are. Much theoretical work is currently being done in this area, but for our purposes it is probably sufficient to consider merely the textbook description of money as a medium of exchange, a unit of account, and a store of value. Money serves these "functions" best because it is the asset which supposedly possesses the greatest degree of "liquidity." Liquidity is often assumed to be composed of two major elements: (*a*) marketability, or a uniform estimate of value across the population, and (*b*) predictability. Although money may be the most marketable asset (it has the smallest bid-ask spread), it is not necessarily the most predictable. It is obviously the most predictable asset in *nominal* terms in that a dollar nearly always sells for close to $1.00, but it is the uncertainty associated with the future *real* value of an asset which has theoretical significance. If future movements of the price level are uncertain, then, unlike the traditional assumption of standard portfolio theory, money cannot be considered a perfectly riskless asset.[11]

Since the uncertainty of the anticipated real value of money is one major element of a money's liquidity and hence can be expected to affect the relationship between the stock and the service flow, a correct estimation of the demand for money requires the inclusion of a variable which measures the predictability of prices. In this paper I use the *variability* of the rate of price change as an operational measure of price unpredictability. This may appear to be a rather crude approximation, but it is actually a good measure if the underlying stochastic structure is assumed to be one of a constant mean plus some random disturbance. Evidence which I have presented in a recent paper (1975) on the autocorrelations of annual rates of price change over various time periods suggests that the assumption of a constant (zero) mean works reasonably well until 1955 and therefore substantiates, at least up to that point, the implicit model of the formation of price expectations used here. Other evidence given in that paper (for example, the short-run impact of changes in the rate of change of prices on the level of interest rates) further suggests that this model of expectations formation did not shift until the 1960s. Variability may therefore be a reasonable approximation to unpredictability for most of our period of analysis.[12]

---

[10] Over the entire 1879–1972 sample period, however, the computed F-test ratio between these two functional forms (using the [0, 1, 1] ARIMA specification of Section III below) is 2.59 ($F_{.05} = 3.95$), and therefore the null hypothesis of the "price" functional form cannot be rejected in favor of the alternative interest rate functional form at the .05 significance level.

[11] Standard portfolio theory is framed in nominal terms, and the only risk is assumed to be the capital value risk associated with unanticipated interest rate changes. Within this analytical framework money is assumed to be a completely safe asset with no uncertainty of return (see Tobin 1958 and Arrow 1971).

---

[12] An editor of this *Journal* suggested that I alternately experiment with explicit models of the formation of price expectations by estimating the time-series process of prices using Box-Jenkins methods. The residuals from this process can then be taken as a measure of unanticipated price change and therefore of price uncertainty. Over the entire 1871–1974 period movements in the price level can be represented by the following (1, 1, 0) ARIMA process:

$$\Delta \log P_t - \underset{(5.21)}{0.4625}\Delta \log P_{t-1} = \underset{(1.67)}{0.00806} + \hat{u}_t$$

$$\bar{R}^2 = .198,\ d = 2.10,\ \mathrm{SE} = 0.0479.$$

The squared residuals from this time-series process did not enter the demand for money significantly. I also divided the period into three subperiods which, as I suggested in my 1975 article, have distinct stochastic properties: (*a*) 1879–1915 turns out to be a random walk in the log of the price level, and therefore the first differences of the series (the actual rate of price change) is white noise or unanticipated price change; (*b*) 1916–55 is a (1, 1, 0) process:

$$\Delta \log P_t - \underset{(3.47)}{0.4611}\Delta \log P_{t-1} = \underset{(1.25)}{0.0137} + \hat{u}_t$$

$$\bar{R}^2 = .150,\ d = 2.22,\ \mathrm{SE} = 0.0655;$$

and (*c*) 1956–72 is a (1, 1, 1) process:

$$\Delta \log P_t - \underset{(4.93)}{0.8068}\Delta \log P_{t-1} = \underset{(1.07)}{0.00577} + \hat{u}_t + \underset{(3.49)}{0.7190}\hat{u}_{t-1}$$

$$\bar{R}^2 = .830,\ d = 1.60,\ \mathrm{SE} = 0.0056.$$

Regressions were also run for overlapping time periods to splice the series of residuals together; however, neither the squared residuals from these processes nor the square root of the sum of the last 6 years' squared residuals divided by five (a measure analogous in form to the variability measure I employ in this paper) entered the demand for money function significantly. I believe this indicates that merely fitting a Box-Jenkins model to past rate of price changes to make price forecasts at every point in time yields misleading results. A major point of my "new standard" paper (1975) is that the public considers other information when forming price expectations, such as the nature of the underlying

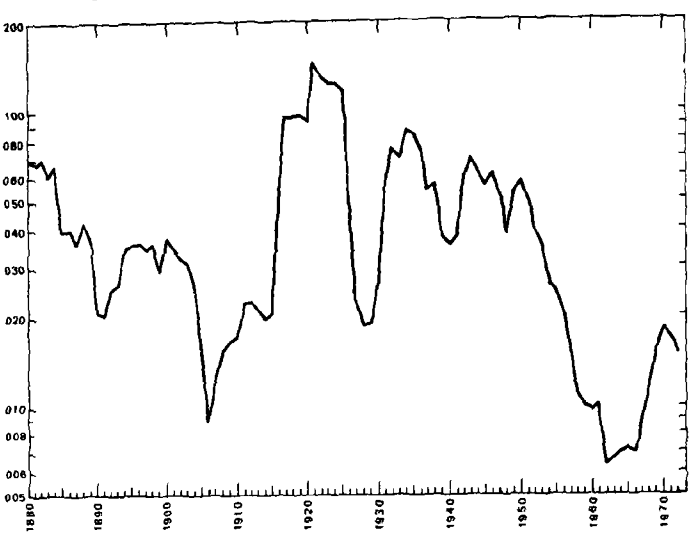


FIG. 2.—Variability of the annual rate of change of prices, 1880–1972

Variability is measured by the five-term moving standard deviation from the 10-term moving mean of the annual rate of change of prices and is plotted in figure 2.[13] This variable, which I label $S(\dot{P}/P)$, is then entered into the U.S. demand for money function. The only previous

---

monetary institutions. The problem is one of explicitly considering shifts in these other factors without relying at any point in time on price information not yet experienced. In addition, I also experimented with measures over time of ex post or realized unpredictability based upon the difference between current market interest rates and actual future rates of price change. But without any independent measure of real rates (e.g., assuming real rate changes were zero), these derived variables were nonsensical. (The 1940s were the most obvious aberration, when interest rates were low while inflation rates were high; interest rates were also very high in 1869–74 while a significant deflation was occurring in the 1870s.) Use of cycle phase average data did not yield more meaningful results.

[13] I computed logarithmic first differences of the implicit national product price deflator and then computed five-term moving standard deviations from these year-to-year percentage rates of price change using as the mean for this standard deviation calculation the average rate of price change over the current and previous 9 years. The result was then dated as of the final year. The vertical scale on the chart is logarithmic to minimize the heteroscedasticity problem. I used a 10- rather than a 5-year mean because I originally thought that the use of a longer-term mean from which the deviations are calculated would produce a better measure of price uncertainty in periods when unusually large (in absolute value) price changes continued for a few years. But the correlation over the 1880–1972 period of this variable with the simple six-term moving standard deviation of the rate of price change that I later used in Klein (1975) was .97, and the demand for money results using this alternative measure are nearly identical. The only other measure of price variability I experimented with was the long-term price uncertainty variable I constructed in Klein (1975) which entered insignificantly over the entire period but was ignificantly positive over the 1880–1918 "gold standard" subperiod.

related empirical work I know of is by Morris Perlman (1970), where the mean squared deviations around the logarithmic trend line of prices were used in estimating the desired currency-to-deposit ratio across countries and also the demand across countries for currency and for deposits. This variable was found in each of these cases to be insignificant.[14]

C. Long-Term Demand for Money Results

Table 1 presents estimates of the demand for money function using annually averaged data. Money and income are defined in real permanent per capita terms, that is, the demand function is assumed homogeneous of the first degree in permanent prices and measured population:[15] r_L is the yield on long-term (up to 30-year) corporate bonds, r_S is the short-term (4–6 month) commercial paper rate, and r_M is the rate of return on money (see Klein 1974*a*). The entire 1880–1972 period is covered for M_2. For the post–World War I 1919–72 period, when reliable data become available for the narrower definition of money, regressions are reported for both M_1 and M_2. In addition, although the strong unchanging trends in the variables over shorter periods cast serious doubt

---

[14] In addition, Brunner and Meltzer (1963) have unsuccessfully used in the demand for money a moving 3-year standard deviation of the level of interest rates as a measure of uncertainty concerning credit market conditions.

[15] Changes in permanent prices and measured population are assumed to cause equal percentage changes in the nominal quantity of money demanded. Some evidence on this proposition with respect to measured prices can be found in Allan Meltzer (1963). Friedman (1969) argues that individuals are likely to determine their desired holdings of money in terms of longer-term price movements, and he successfully uses a permanent price index deflator. But Friedman's series, which is merely a weighted average of past measured price levels with no adjustment made for trend, cannot be meaningfully extended through the 1960s. During that period a dollar fiduciary standard with a clear positive long-term trend in the price level replaced any remaining semblance of a gold commodity standard with its general presumption of a long-term stable price level (see Klein 1975). (The downward bias is obvious when one considers the level of the ratio of measured to unadjusted permanent prices, which is greater than one every year since 1940, is 1.03 in 1963 before the rapid inflation begins, and reaches a level of 1.08 in 1972.) Any attempt to adjust the permanent price level series must be somewhat arbitrary. I use the rising level of the long-term interest rate as a crude market measure of the postwar acceleration of the long-term trend of prices, and I measure trend over the 1952–72 period by subtracting the long-term rate of interest from the average long-term rate during 1946–51, smoothing a monotonically increasing series to the nearest quarter of a percentage point. The adjusted permanent price level series remains below measured prices, but the measured to permanent price ratio is only 1.01 in 1963 and rises to only 1.03 in 1972. Also, the deflation by measured population (the use of per capita averages of aggregated income and money holdings) may not produce an appropriate approximation. Some measure of the dispersion of income and money holdings may also be relevant. In addition, since foreign holdings of currency are included in the money estimates, the U.S. population figures used are somewhat arbitrary. All permanent estimates are made using Friedman's (1957) consumption function weights, but see n. 24 below for an empirical justification of this ad hoc procedure.

TABLE 1

DEMAND FOR MONEY FUNCTIONS, ANNUAL OBSERVATIONS, 1880–1972 AND 1919–72

$$\log M = a_0 + a_1 \log y_p + a_2 r_S + a_3 r_L + a_4 r_M + a_5 \log S(\dot{P}/P)$$

Dependent Variable and Time Period	a_0	a_1	a_2	a_3	a_4	a_5	R^2	d	SE
M_2 1880–1972:									
(9)....... ..	−13.71	1.307	−0.299	−0.057	0.324	...	.989	1.07	0.0755
	(65.69)	(40.73)	(12.06)	(4.39)	(10.96)				
(10)...	−14.01	1.380	−0.268	−0.071	0.302	0.050	.991	0.98	0.0684
	(69.55)	(41.44)	(11.37)	(5.86)	(11.09)	(4.48)			
M_2 1919–72:									
(11)... .. .	−13.40	1.251	−0.198	−0.032	0.187	...	.961	0.46	0.0724
	(50.16)	(31.63)	(2.86)	(2.13)	(2.32)				
(12)........	−14.38	1.433	−0.180	−0.036	0.173	0.074	.976	0.73	0.0579
	(50.94)	(30.79)	(3.23)	(2.98)	(2.67)	(5.34)			
M_1 1919–72:									
(13)........	−13.95	1.322	−0.170	−0.125	0.176	...	.918	0.36	0.1167
	(34.77)	(21.94)	(1.62)	(4.63)	(1.35)				
(14)........	−15.80	1.668	−0.135	−0.135	0.146	0.143	.962	0.75	0.0806
	(42.29)	(26.65)	(1.85)	(7.21)	(1.63)	(7.40)			

NOTE.—All logs stand for natural logarithms. The absolute values of the t-statistics are given in parentheses beneath the coefficient estimates, d represents the D-W statistic; R^2, the coefficient of determination.

on the effective number of degrees of freedom present within smaller subperiods, I also examine the post–World War II period separately.

All of the variables enter in the expected direction and, except for r_S and r_M in the 1919–72 regression, all of the reported t-statistics are greater in absolute value than 2.[16] Note, in particular, that the variable measuring the uncertainty of price change positively enters the demand for money. Given price inelasticity of demand in the region under consideration, the positive effect of price variability on the demand for money is a result predicted by the theory.

The r_M variable is the same as used in my earlier article (1974*a*), except for two adjustments: (1) Commercial bank float was mistakenly not subtracted out of my estimate of reserves held (at a cost) by commercial banks against deposits for the 1968–70 period. This mistake is now corrected.[17] (2) Service charges on demand deposits are no longer

[16] Since serious first-order positive serial correlation of the error terms is indicated by the reported Durbin-Watson statistics the t-test is invalid, and statements about statistical significance therefore cannot be made. In the next section of the paper the autocorrelation of the residuals is analyzed, and this problem is corrected.

[17] This mistake was made because the Federal Reserve's somewhat confusing treatment of bank float was not recognized. In computing *private demand deposits at member banks* (and therefore money stock measures), the Fed subtracts from gross deposits cash items in the process of collection at member banks, while Federal Research float is *not* subtracted when computing their *member bank deposit* series. The rationale for this seemingly inconsistent procedure is that the former series is designed to measure "the money balances of the nonbank sector" while the latter series is supposed to measure "the deposit resources of member banks" (see *Federal Reserve Bulletin* [February 1973], pp. 62–66).

subtracted out in computing r_M. Since service charges enter on a per unit time and/or per transaction basis, rather than on a per dollar deposit basis, I did note (Klein 1974*a*, n. 17) that considering service charges as negative interest on money made no theoretical sense. I nonetheless (mistakenly) subtracted them when constructing r_M because such a procedure "worked better" empirically. When the price uncertainty variable is added to the money demand equation, however, that procedure produces an inferior fit. Once the postwar fall in price variability is considered (see fig. 2), it is no longer necessary to use the rising postwar service charge rate on deposits to "explain" (incorrectly) the postwar rise in velocity.

D. Velocity Results

The demand for money results of table 1 are replicated in table 2 using velocity as the dependent variable.[18] Once again all variables have reported t-statistics greater in absolute value than 2, except r_S and r_M in the 1919-72 regressions, but there is once again a great deal of positive serial correlation of the residuals. Log $S(\dot{P}/P)$ enters in the theoretically correct negative direction and moves the coefficient on log (Y/Y_p) (the ratio of measured to permanent nominal income) closer in the 1919–72 period to the value of 1, which would be theoretically appropriate under the assumption that transitory income does not affect the long-run demand for money.

Perhaps the most dramatic results with the price uncertainty variable occur over the postwar period. The velocity regression results for the 1946-72 and 1953-72 subperiods reported in table 3 indicate a marked improvement in the fit of the regressions when the price uncertainty variable is added. The standard error of estimate of the regression is reduced by 55 percent for 1946–72 and by nearly 80 percent for 1953–72. The estimated coefficients also move toward theoretically appropriate values. (The implied income elasticity moves from $-.8$ to $+.4$ and from $-.9$ to 0, the long-term interest rate effect moves from zero or slightly negative to positive, and the (Y/Y_p) coefficient moves from zero or slightly negative to the theoretically expected positive, or procyclical, direction.)[19]

---

[18] Velocity is defined as nominal net national product divided by the money stock. This alternative somewhat more general specification differs from the initial demand for money specification in not using population as a deflator and in failing to constrain the (Y/Y_p) coefficient to 1. Kuznets's income figures were spliced to the post-1946 Department of Commerce figures by multiplying the department's figures by 0.882—the ratio of Kuznets's 1946 NNP estimate to the department's 1946 estimate.

[19] The coefficient on log (Y/Y_p) of greater than 1 for 1946–72, however, is theoretically disturbing since it implies that an increase in transitory income leads to a decrease in money holdings. The coefficient is not, however, statistically significantly different from 1, and if there are measurement errors in nominal income this coefficient will be biased upward due to spurious correlation.

TABLE 2

VELOCITY FUNCTIONS, ANNUAL OBSERVATIONS, 1880–1972 AND 1919–72

$$\log V = a_0 + a_1 \log y_p + a_2 r_S + a_3 r_L + a_4 r_M + a_5 \log (Y/Y_p) + a_6 \log S(\dot{P}/P)$$

Dependent Variable and Time Period	a_0	a_1	a_2	a_3	a_4	a_5	a_6	R^2	d	SE
V_2 1880–1972:										
(15)	2.177	−0.312	0.289	0.065	−0.318	1.059	. . .	.948	1.03	0.0725
	(10.85)	(10.05)	(12.11)	(4.92)	(11.16)	(13.29)				
(16)	2.489	−0.389	0.256	0.082	−0.298	1.132	−0.052	.959	0.92	0.0651
	(12.97)	(12.03)	(11.37)	(6.71)	(11.47)	(15.48)	(4.70)			
V_2 1919–72:										
(17)	1.869	−0.242	0.065	0.021	−0.026	0.726	. . .	.784	0.48	0.0673
	(7.49)	(6.36)	(0.75)	(1.30)	(0.25)	(5.55)				
(18)	2.869	−0.436	0.158	0.039	−0.146	0.967	−0.072	.858	0.73	0.0551
	(9.98)	(8.70)	(2.11)	(2.87)	(1.64)	(8.21)	(4.95)			
V_1 1919–72:										
(19)	1.592	−0.181	−0.162	0.115	0.264	0.205	. . .	.813	0.53	0.1001
	(3.94)	(2.85)	(1.34)	(4.89)	(1.70)	(1.06)				
(20)	3.624	−0.554	−0.024	0.131	0.064	0.611	−0.122	.897	0.76	0.0749
	(8.15)	(7.24)	(0.26)	(7.38)	(0.53)	(3.84)	(6.23)			

NOTE.—See note to table 1.

TABLE 3

VELOCITY FUNCTIONS, ANNUAL OBSERVATIONS, 1946–72 AND 1953–72

$$\log V = a_0 + a_1 \log y_p + a_2 r_L + a_3 \log (Y/Y_p) + a_4 \log S(\dot{P}/P)$$

Dependent Variable and Time Period	a_0	a_1	a_2	a_3	a_4	R^2	d	SE
V_1 1946–72:								
(21)	−11.555	1.782	−0.009	0.040	. . .	.858	0.18	0.0974
	(3.22)	(3.29)	(0.18)	(0.05)				
(22) . . .	−3.957	0.558	0.058	1.496	−0.150	.972	0.90	0.0441
	(2.18)	(2.01)	(2.55)	(4.03)	(9.48)			
V_1 1953–72:								
(23)	−12 460	1.949	−0.048	−1.183	. .	.913	0.66	0.0535
	(4.18)	(4.33)	(1.35)	(1.42)				
(24) . . .	−6.714	1.006	0.020	0.558	−0.112	.996	1.83	0.0120
	(9.01)	(8.79)	(2.28)	(2.63)	(17.44)			

NOTE. *See note to table 1*

In addition, the Durbin-Watson statistic improves dramatically, with the null hypothesis of zero first-order serial correlation in the error terms accepted at the .05 significance level for the 1953–72 regression.

These postwar regressions are reported without r_S and r_{M_1} because neither enters the relationship significantly. In fact, before the price uncertainty variable is added, r_S enters the relationship in the theoretically incorrect negative direction ("reported" t-statistic equal to −2.8). This negative coefficient is picked up in the data because the rate of growth of V_1 flattens out in 1967–71 just when short-term interest rates begin to rise. When the price uncertainty measure is introduced, the 1967–71 slowdown in the growth of V_1 is now "explained" by an increase in price uncertainty rather than by a rise in interest rates. Short-term rates then enter positively, but not significantly. In addition, since the annual rate of change in velocity is extremely stable during the postwar period, the elimination of r_S and r_{M_1} can also be rationalized as necessary to create additional degrees of freedom for statistically meaningful results.

The improvement in the empirical results over the postwar period can perhaps be illustrated most clearly by considering figures 3 and 4. Figure 3, which plots the values of log V_1 predicted by equation 21 (along with the actual values), reveals that this equation picks up little of the movement over time in velocity other than trend. In particular, the unpredicted slowdown in velocity growth after 1967 is contrary to what would be expected from interest rate changes over this period.[20] Conse-

---

[20] If competitive interest is paid on deposits, an increase in the anticipated inflation rate and therefore market interest rates will also increase r_M and hence theoretically suggest a small total measured effect on the demand for money. My earlier results (Klein 1974*a*, pp. 944–45) implied that, evaluated at sample means over the 1880–1970 period, a 1 percentage point increase in the expected inflation rate would decrease the

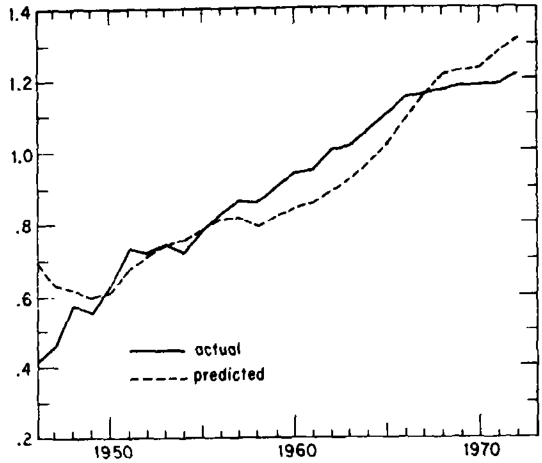


FIG. 3.—Actual and predicted values of V_1 from regression equation (21), 1946–72

quently, the residuals from this regression, plotted in figure 4 (*residual*$_1$), become quite negative at the end of the period when predicted velocity persistently overestimates actual velocity. When the price uncertainty term is added, however (eq. [22]), the residual (plotted in fig. 4 as *residual*$_2$) is greatly reduced overall, and it essentially vanishes during the difficult post-1967 period. The movement in price uncertainty may therefore explain much of the unaccounted-for movement in velocity since 1967 that previously has been a mystery.[21]

III. Further Empirical Results

Almost all the regression results so far reported have extremely low D-W statistics, indicating a likely misspecification of the relationship. I therefore attempt in this section to correct this misspecification by examining (*a*) the time-series process on the residuals, (*b*) the functional form of the equation, and (*c*) the possibility that a relevant variable has been omitted.

---

demand for money only about 0.1 percent. But it certainly should not increase the demand for money. Cagan and Schwartz (1975) have also recently noted their inability to explain the sharp break in the trend rate of growth of velocity that occurs during this period.

[21] The postwar results are so striking that the significant effect of log $S(\dot{P}/P)$ over the entire 1880–1972 period may be thought to be due solely to the movement in this subperiod. This is, however, not the case. Log $S(\dot{P}/P)$ enters positively over the 1880–1945 subperiod with a reported t-statistic greater than 2. So the results, although weaker in the earlier period, do not indicate a purely postwar effect. The pre-1946 national income deflator is a relatively poorer price index in terms of coverage and accuracy. This measurement problem adds noise to the price uncertainty variable in the early period which would be expected to reduce its significance.

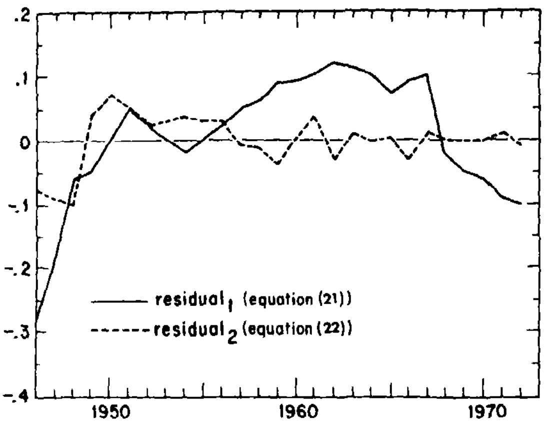


FIG. 4.- V_1 regression equation residuals (actual minus predicted values) equations (21) and (22), 1946–72.

A. Time-Series Analysis[22]

The demand for money regressions in the levels are clearly not stationary; therefore, the equations are presented in table 4 in first differenced form with a moving average error term for the M_2 regressions and an autoregressive error term for the M_1 regression (the ARIMA processes which best eliminated the autocorrelation in the residuals). The Q-statistics for the residuals (the Box-Pierce statistic adjusted for downward bias, an approximate χ^2 test of the randomness of the first 12 sample autocorrelations) are reported, and they are all small relative to the critical values at the .05 significance level.

The results are quite reassuring. The variables generally enter significantly in their theoretically expected directions after correcting for the extreme positive serial correlation present in the original regressions. In particular, the price uncertainty variable enters both positively and significantly (except for 1919–72 M_2 where it is barely insignificant at the .05 level).[23] In addition, it may be worth noting that the estimated permanent income elasticity of demand for M_2 is now almost exactly 1, a value that makes theoretical sense if money is considered an input

---

[22] I am grateful to Tom Heagy for assistance on this section. The results reported here are based upon the more complete tests conducted in his Ph.D. dissertation currently in progress at the University of Chicago Graduate School of Business.

[23] After examining the cross-correlations (the correlations of the independent variables with the residuals) it became apparent that the price uncertainty variable should enter in lagged form. Since this also makes a great deal of theoretical sense (otherwise individuals are basing their money holdings on price variability observations that have not yet occurred), that is how the variable is entered and reported in table 4.

TABLE 4

DEMAND FOR MONEY FUNCTIONS, ANNUAL OBSERVATIONS, 1879–1972 AND 1919–72

$$\Delta \log M_2 = a_0 + a_1 \Delta \log y_p + a_2 \Delta r_S + a_3 \Delta r_L + a_4 \Delta r_{M_2} + a_5 \Delta \log S(\dot{P}/P)_{t-1} + \hat{u}_t - \theta \hat{u}_{t-1}$$

Time Period	a_0	a_1	a_2	a_3	a_4	a_5	θ	$\bar{R}^2$	d	Q	SE
1879–1972: (25)	0.0109 (2.19)	1.009 (6.90)	−0.0508 (3.41)	−0.0604 (4.92)	0.0671 (3.14)	0.0234 (2.11)	−0.3901 (3.71)	.671	1.98	8.01	0.0299
1919–72: (26)	0.0035	1.031 (6.10)	−0.964 (3.33)	−0.0437 (3.37)	0.1183 (3.27)	0.0273 (1.70)	−0.4881 (3.63)	.710	1.94	4.34	0.0284

$$\Delta \log M_1 = a_0 + a_1 \Delta \log y_p + a_2 \Delta r_S + a_3 \Delta r_L + a_4 \Delta r_{M_1} + a_5 \Delta \log S(\dot{P}/P)_{t-1} + \hat{u}_t/(1 - \phi B)$$

Time Period	a_0	a_1	a_2	a_3	a_4	a_5	ϕ	$\bar{R}^2$	d	Q	SE
1919–72: (27)	0.0035	0.8702 (3.94)	−0.0887 (3.00)	−0.0258 (1.75)	0.1180 (2.92)	0.0110 (2.49)	0.6667 (6.13)	.749	1.82	7.02	0.0316

NOTE.—See note to table 1, and B is the backshift operator.

entering household and firm production functions which are homogeneous of the first degree in all inputs. The results differ from those of previous long-run demand for money studies by indicating not only that money is not a luxury good, but also that the income elasticity of demand for money does not seem to fall over time. There were not enough degrees of freedom to employ meaningfully the multiple input transfer function program used for this time series estimation on the postwar data alone.[24]

B. Multiplicative Functional Form

If the demand for money function is semilogarithmic in the interest rates, as assumed in these empirical estimates, and rotates around the point of unitary elasticity, then the price uncertainty variable will enter the demand for money as a function of the price of money; therefore, the multiplicative functional form, $(r_S - r_M) \cdot \log S(\dot{P}/P)$, might be expected to yield superior results. These multiplicative estimates are presented in table 5 and are generally superior. The standard error of estimate is reduced about 7 percent for the M_2 regressions while essentially remaining unchanged for the M_1 regression. In addition, the price uncertainty variable in the 1919–72 regression now enters positively at the usual levels of significance, and in the 1879–1972 regression it enters substantially more significantly.

C. Real Income Uncertainty

Although the price uncertainty variable enters the demand for money significantly, it is possible that it is merely a proxy for the degree of general economic instability or real income uncertainty in the economy. This omitted variable may be expected to affect the demand for money by directly influencing the demand for liquidity.[25] The demand for

---

[24] The multiple input transfer function (MITF) model is discussed by Box and Jenkins. The computer program used here was developed by Charles Nelson (TRANSEST, University of Chicago Graduate School of Business, August 1975). It is interesting to note that when actual rather than permanent real per capita income was entered into this demand for money model with a single denominator term, i.e., a Koyck distributed lag structure, the estimated lag weights coincided exactly with Friedman's earlier estimates (0.33 on the current value), although Friedman's estimates were derived for a completely different (consumption) model from a different body of data over a different time period.

[25] Fisher (1932) claimed that pessimism or a general loss of confidence in business would increase the demand for money. A somewhat similar effect may have been operating at the outbreak of war in 1914 and again in 1939 when an increase in general uncertainty in the economy produced a decrease in velocity. Although war is associated with rapid inflation and therefore one would expect a movement out of money, the initial movement was a decrease in velocity. The large amount of real uncertainty that must have been present in the economy immediately after World War II may also possibly explain the large negative velocity residuals during this period (see fig. 3). In

TABLE 5

Demand for Money Functions, Annual Observations, 1879–1972 and 1919–72

$$\Delta \log M_2 = a_0 + a_1 \Delta \log y_p + a_2 \Delta r_S + a_3 \Delta r_L + a_4 \Delta r_{M_2} + a_5 \Delta[(r_S - r_{M_2}) \log S(\dot{P}/P)]_{t-1} + \hat{u}_t - \theta \hat{u}_{t-1}$$

Time Period	a_0	a_1	a_2	a_3	a_4	a_5	θ	$\bar{R}^2$	d	Q	SE
1879–1972: (28)	0.0097 (2.07)	0.9501 (6.94)	−0.0592 (4.21)	−0.0494 (3.87)	0.0712 (3.53)	0.0072 (4.23)	−0.4015 (3.87)	.714	1.96	7.06	0.0278
1919–72: (29)	0.0033 (0.52)	0.9589 (5.99)	−0.1168 (4.28)	−0.0297 (2.54)	0.1346 (4.02)	0.0172 (3.23)	−0.5141 (3.99)	.749	1.84	4.63	0.0264

$$\Delta \log M_1 = a_0 + a_1 \Delta \log y_p + a_2 \Delta r_S + a_3 \Delta r_L + a_4 r_{M_1} + a_5[(r_S - r_{M1}) \log S(\dot{P}/P)]_{t-1} + \hat{u}_t/(1 - \phi B)$$

Time Period	a_0	a_1	a_2	a_3	a_4	a_5	ϕ	$\bar{R}^2$	d	Q	SE
1919–72: (30)	0.0055 (0.44)	0.9236 (4.26)	−0.0869 (2.87)	−0.0444 (3.08)	0.1247 (3.00)	0.0420 (2.39)	0.6127 (5.50)	.746	1.88	8.13	0.0317

money regressions were therefore repeated, including an additional variable to measure the degree of stability of the economy analogous to the price uncertainty measure—a moving five-term standard deviation from a 10-term mean of the rate of change in real per capita income, $S(\dot{y}/y)$.[26] The results for the 1879–1972 period are[27]

$$\begin{aligned}
\Delta \log M_2 = {} & \underset{(1.68)}{0.0080} + \underset{(8.24)}{1.060\Delta \log y_p} - \underset{(3.19)}{0.0345\Delta r_L} - \underset{(4.67)}{0.0419\Delta(r_S - r_{M_2})} \\
& + \underset{(4.50)}{0.0076\Delta[(r_S - r_{M_2}) \log S(\dot{P}/P)]_{t-1}} \\
& + \underset{(0.77)}{0.0085\Delta \log S(\dot{y}/y)_{t-1}} + \hat{u}_t + \underset{(4.28)}{0.4242\hat{u}_{t-1}}
\end{aligned} \tag{31}$$

$$\bar{R}^2 = .706,\ d = 1.96,\ Q = 9.40,\ \mathrm{SE} = 0.0282.$$

Real income variability does not significantly enter the demand function, and therefore the final equations of table 5 are not misspecified in this regard.[28]

---

addition, the real uncertainty associated with *relative* price unpredictability can be assumed theoretically to enter the demand for money in a positive direction. There is no a priori reason, however, to believe that relative price uncertainty is positively correlated with my measure of absolute price uncertainty.

[26] The simple correlation of this variable with the price uncertainty variable is .63. Since real income is measured residually from nominal income and prices, there is the possibility of measurement errors producing spurious correlation between the two series. I also used over the 1919–72 period the spread between Moody's Baa and Aaa corporate bond rates as a possible market measure of real uncertainty, but this variable did not enter significantly.

[27] The regression must be run in the price functional form because Nelson's transfer function computer program can only accommodate five independent variables. The results, however, are likely to be very similar to the simple interest rate form (see n. 10).

[28] Another possible source of misspecification concerns the form of the rate of return on money variable. In a recent paper, Eden (1976) demonstrates that the usual specification of the expected rate of return on money as the negative of the expected inflation rate, $-(\dot{P}/P)^*$, is incorrect. He shows that since individuals are interested in the exchange value of money, $1/P$, which is a convex function in $P > 0$, the expected rate of return on money is more precisely $\{1/[1 + (\dot{P}/P)^*]\} - 1$ and that this correct expected rate of return increases with the even moments and decreases with the odd moments of the continuous rate of inflation. In particular, this correct expected real rate of return increases with the variance and decreases with the mean of the rate of inflation. Using data from hyperinflation episodes, Eden then shows that a measure of the anticipated variance generally has a significant positive effect on the demand for money only when the rate of inflation rather than the correct return on money is used. It is difficult, however, to apply these results in this context to my equations. Whether $E[1/(\dot{P}/P)]$ or $-E(\dot{P}/P)$ is used will make very little difference except during periods of very rapid price changes. The two alternative measures are obviously very close for this sample, and clearly the price variability term is not picking up solely this effect.

IV. Implications

These empirical results have important theoretical implications since they suggest that an increase in monetary instability which produces an increase in the variability of prices would lead individuals to voluntarily hold additional cash balances. If the traditional conceptual framework of inflation analysis is employed, the primary distinction between the effects of an "anticipated" and an "unanticipated" inflation is now blurred. It is commonly maintained that only an anticipated inflation will affect the demand for money, with unanticipated inflation merely producing once-and-for-all wealth distribution effects. The empirical results of this paper, however, indicate that unanticipated price changes, by altering the degree of future price uncertainty (or the variance of the prior probability of future price change distribution), will now also have allocative effects. In addition to any distributional effects, an unanticipated price change will also alter the future demand for money, but in the opposite direction to an anticipated price change.

Within the context of the optimum quantity of money literature, this increase in the demand for (and therefore the level of) real money that is produced by increased price uncertainty does *not* imply an increase in social welfare. A possible source of error here is a confusion between the stock and the service flow. Utility is not a function of the stock of real cash balances, (M/P), but of the monetary service flow. An increase in price uncertainty will reduce the actual monetary service stream as long as $N_{(M/P)}$ falls and therefore the price of monetary services rises. Society is clearly worse off because a valuable asset (price predictability) has been depreciated.[29]

---

[29] In the same optimum quantity of money context it is also important to distinguish clearly between the stock and the service flow when considering the marginal costs of producing money. The usual optimum solution has a rate of deflation equal to "the" real rate of interest so that the price of holding money is set equal to the assumed zero marginal cost of producing money. But in what sense is money produced costlessly on the margin? The fact that it is costless to "add a zero," i.e., that the marginal cost of nominal money is zero, implies nothing about the marginal cost of producing monetary services, or real cash balances, which is the relevant concept here. It is difficult to conceptualize "the marginal cost of producing real cash balances" since (M/P) is not a control variable in a money-producing firm's production function but rather is determined on the demand side. A useful way to think of this theoretically is to start with the public's demand for monetary services, which implies a derived demand for quality or monetary confidence, β (similar to the derived demand for entrepreneurial services). Given this demand for β, the marginal cost of producing β (for example, holding gold stocks, maintaining an army and police force, or expending resources on any other inputs which create confidence capital) implies an equilibrium quantity of β and thus a demand for real cash balances function. More generally, given information costs about future performance, monetary confidence or "brand names" will have value independent of any confidence-creating expenditures. Even if the costs of "producing" real cash balances are zero, consumer demand for collateral or "backing" would necessarily imply a positive price such that a normal rate of return is earned on the residually measured intangible brand name capital. This "profit" can be thought of as the cost of "selling" the real cash balances (see Klein [1974*b*] for a related discussion).

What has increased, however, is the potential governmental tax revenue from inflation. An unanticipated inflation will increase price uncertainty and therefore will, in general, increase the tax base. What places a limit on this induced growth of the base via the creation of price uncertainty is the degree of competition (or the implied elasticity of demand for monetary services) present in the particular market for money. All real world examples lie somewhere between the perfectly competitive market for monetary services of my earlier theoretical model (1974*b*), where the marginal productivity or complementarity-in-consumption effect must dominate,[30] and the monopolistic case empirically investigated in this paper, where the average productivity or substitution-in-production effect has clearly dominated. The direction of the effect of price uncertainty changes on money demand will depend upon the particular market and time horizon under consideration. One would expect the elasticity of demand to be greater in the long run than in the short run and to be greater in the relatively more competitive market for international money compared to a particular monopolistic money market. Eventually a policy of creating price uncertainty, which "depreciates" the money's "brand name," will very likely cause an erosion of the tax base, but the empirical results of this paper have demonstrated that this "competitive" effect may be quite sluggish.[31]

References

Arrow, K. "The Theory of Risk Aversion." In *Essays in the Theory of Risk Bearing.* Chicago: Markham, 1971.

Box, G. E. P., and Jenkins, G. M. *Time Series Analysis.* San Francisco: Holden-Day, 1970.

Brunner, K., and Meltzer, A. H. "Some Further Investigations of Demand and Supply Functions for Money." *J. Finance* 18 (May 1963): 319–54.

Cagan, P., and Schwartz, A. J. "Has Growth of Money Substitutes Hindered Monetary Policy?" *J. Money, Credit and Banking* 7 (May 1975): 137–59.

Eden, B. "On the Specification of the Demand for Money: The Real Rate of Return versus the Rate of Inflation." *J.P.E.* 84, no. 6 (December 1976): 1353–59.

Fisher, I. *Booms and Depressions.* New York: Adelphi, 1932.

---

[30] In a perfectly competitive market for monetary services, a fall in the marginal productivity of a firm's money, no matter how small, raises the real implicit price of monetary services from the firm's real cash balances and drives the firm's demand to zero.

[31] In the international money market this effect, on the contrary, may be quite rapid and large. See Klein (1976) for a further discussion of these issues. It is important to recognize the theoretical forces at work here. The recent rise in the market price of gold, for example, has significantly reduced gold's liquidity or monetary service flow and should ultimately decrease the demand for gold relative to competing international monies such as the dollar. Although holders of gold have experienced a large capital gain over the last decade, the existence of volatile (unpredictable) price movements suggests that gold has merely become a speculative commodity similar to, say, soybeans or cocoa and is no longer suitable as a money.

Friedman, M. "The Quantity Theory of Money—a Restatement." In *Studies in the Quantity Theory of Money*. Chicago: University of Chicago Press, 1956.

———. *A Theory of the Consumption Function*. Princeton, N.J.: Princeton Univ. Press (for Nat. Bur. Econ. Res.), 1957.

———. "The Demand for Money—Some Theoretical and Empirical Results." In *The Optimum Quantity of Money and Other Essays*. Chicago: Aldine, 1969.

Harberger, A. *The Demand for Durable Goods*. Chicago: Univ. Chicago Press, 1960.

Klein, B. "Competitive Interest Payments on Bank Deposits and the Long-Run Demand for Money." *A.E.R.* 64, no. 5 (December 1974): 931–49 (*a*).

———. "The Competitive Supply of Money." *J. Money, Credit and Banking* 6, no. 4 (November 1974): 423–53. (*b*).

———. "Our New Monetary Standard: The Measurement and Effects of Price Uncertainty 1880–1973." *Econ. Inquiry* 13, no. 4 (December 1975): 461–84.

———. "Competing Monies: A Comment." *J. Money, Credit and Banking* 8, no. 4 (November 1976): 513–19.

Meltzer, A. H. "The Demand for Money: The Evidence from the Time Series." *J.P.E.* 71, no. 3 (June 1963): 219–246.

Perlman, M. "International Differences in Liquid Assets Portfolios." In *Varieties of Monetary Experience*, edited by D. Meiselman. Chicago: Univ. Chicago Press, 1970.

Tobin, J. "Liquidity Preference as Behavior towards Risk." *Rev. Econ. Studies* 25 (February 1958): 65–86.

The Competitive Effects of Trust-Busting: A Portfolio Analysis

---

Malcolm R. Burns

University of Kansas

Monthly stock prices are used to analyze the competitive effects of the Sherman Act dissolutions carried out against Standard Oil, American Tobacco, and American Snuff on December 1, 1911. Statistical tests of the price changes accompanying important developments in the litigation show rather strongly that investors did not expect dissolution to alter the performance of the snuff, tobacco, and petroleum industries. This conclusion is similar to the findings of previous studies of these cases, and the results indicate that applications of stock prices and finance theory can significantly augment the empirical research methods of industrial organization.

I. Introduction

Recent empirical work in corporate finance supports both the capital asset pricing model (CAPM) and the efficient markets hypothesis of stock price behavior.[1] For example, the time series estimates of Fama and MacBeth (1973) have verified the predicted linear and risk-adjusted relationship between individual portfolio returns and the market average.

This article is a condensation of my Ph.D. dissertation in economics at the University of Minnesota. The research was supported by grants from the Brookings Institution and the University of Kansas. I would like to thank Uri Ben-Zion, Peter Clark, Mohamed El-Hodiri, Anthony Marino, Trout Rader, and Leo Raskind for helpful comments and suggestions on an earlier draft. I am especially grateful to John Hause, who supervised this project as chairman of my Ph.D. committee. The revision has benefited substantially from the criticisms of an anonymous referee and the assistance of Trent Boggess, Maurice Joy, and De-Min Wu. I am responsible for any remaining errors. A preliminary version of this paper was presented at the Kansas-Missouri Quantitative Economics Seminar on May 2, 1975.

[1] The CAPM was developed initially by Tobin (1958) and Markowitz (1959). It is discussed at length in Fama and Miller (1972, pp. 215–320). A comprehensive survey of the efficient markets hypothesis is contained in Fama (1970).

[*Journal of Political Economy*, 1977, vol. 85, no. 4]
© 1977 by The University of Chicago. All rights reserved.

Likewise, studies of the price revisions accompanying stock splits, earnings declarations, and changes in accounting practices indicate that security prices adjust promptly and accurately to new information.[2] So far the research has concentrated on testing the validity of these models, but the accumulated evidence is now sufficiently strong that financial analysis can be profitably applied to other problems in economics. Thus, instead of determining that security prices react as expected to events whose consequences are largely known ex ante, their behavior can be used to identify the meaning and significance of the events themselves.

Industrial organization is a fertile field for such applications because the monthly character of stock prices makes them an effective substitute for annual measures of market performance which are often unavailable.[3] One topic that has generated new interest is the economic impact of trust-busting, and the experience with previous Sherman Act cases should provide a useful historical framework for evaluating current efforts (e.g., AT & T and IBM) to foster rivalry through dissolution. Consequently this study reexamines the competitive effects of the horizontal dissolutions carried out against Standard Oil, American Tobacco, and its subsidiary American Snuff on December 1, 1911.[4] The sample is small but surprisingly diverse, and the principal features of each decree are outlined in table 1.[5]

Security prices reveal the impact of dissolution quite dramatically. At a minimum, these data indicate that the performance of the snuff, tobacco, and petroleum industries was not significantly altered by the replacement of each trust with independent rivals. However, if one accepts the prevailing view that the three defendants were empirical examples of textbook monopolies, the results also imply that there was very little actual or anticipated competition between the firms in table 1 and no visible erosion of monopolistic rents.[6] This particular finding is identical to the conclusions of the dissolution literature, but it emerges without many of the ambiguities that characterize industry studies.[7] The use of security prices also clarifies what could be an inescapable dilemma of section 2 enforcement: balancing the consumer interest in vigorous rivalry with shareowner concern for the preservation of monopoly profits and capital values. Stockholder injury is an important consideration in dissolution proceedings, and the infrequent use of structural remedies is usually blamed on the reluctance of the courts to punish allegedly "innocent" investors.[8] In short, the results suggest that financial analysis

---

[2] See Ball (1972), Fama et al. (1969), and Jordan (1973).

[3] Previous uses of security prices in industrial organization include Ellert (1975), Lev and Mandelker (1972), Ornstein (1972), and Stigler (1965). Ellert's study is especially noteworthy because he used similar methods to analyze the relative price behavior of 1,466 common stocks issued by corporations involved in 1,019 antitrust cases between 1935 and 1971. Ellert also concludes (pp. 127, 129) that security prices are generally sensitive to new information with antitrust implications.

[4] *Standard Oil Company of N.J. v. U.S.*, 221 U.S. 1 (1911) and *U.S. v. American Tobacco Co.*, 221 U.S. 106 (1911). The complete texts of the dissolution decrees are presented in U.S. Courts (1918, pp. 136–44, 162–91). The snuff-tobacco decree is summarized in both the October 21, 1911 issue of the *Commercial and Financial Chronicle* (*CFC*), pp. 1122–24, and *Poor's Manual of Industrials*, 1912, pp. 1643 44, 1657–60.

[5] According to Whitney (1958, 2: 386–87), the Justice Department filed 17 Sherman Act suits against large manufacturing corporations between 1900 and 1920. The sample was restricted to the snuff, tobacco, and petroleum trusts primarily because of the expense of collecting substantial amounts of data from original sources. Studies of financial events in more recent periods typically avoid this problem with the use of the stock price tapes compiled by the Center for Research in Security Prices (CRSP) at the University of Chicago; this series begins in January 1926 as described in Fisher and Lorie (1964, pp. 2-3). Of the remaining cases in Whitney, only *U.S. v. E. I. du Pont de Nemours & Co.*, 188 Fed. 127 (D. Del. 1911) resulted in dissolution, and it was excluded due to insufficient stock price data. The other suits produced acquittals or minor consent decrees. Thus, the sample covers all but one of the early manufacturing dissolutions, including the landmark cases of *Standard Oil* and *American Tobacco*, but the empirical results are not necessarily descriptive of all antitrust suits filed during the period 1900–1920.

---

[6] Unfortunately, as is often the case in industrial organization, the empirical results will support other inferences besides "no competition between the successor companies." For example, it is conceivable that the histories of the snuff, tobacco, and petroleum trusts contained in U.S. Bureau of Corporations (1907 and 1909–15) overestimated both their market power and the extent of output restriction. If so, then the absence of negative effects from dissolution in the stock price data suggests that these industries were essentially competitive prior to 1912. Williamson and Andreano (1965) argue that such misconceptions are prevalent in discussions of Standard Oil. The available evidence for all three defendants is reviewed in Burns (1975, pp. 12–36), and some of it is consistent with the conventional view. Thus the defendants controlled substantial shares of most industry branches:

American Snuff (1900–1910)		American Tobacco (1900–1910)		Standard Oil (1899–1911)	
Snuff	91%	Cigarettes	85%	Kerosene	80%
		Little cigars	79	Gasoline	75
		Chewing	77	Waxes	59
		Fine-cut	74	Fuel oil	58
		Smoking	69	Lubricants	47
		Cigars	13		

U.S. Bureau of Corporations (1915, p. 2) and Williamson and Andreano (1965, p. 309). Likewise, the Bureau of Corporations reported that the purchase of American Tobacco common shares in 1890 would have generated income and capital gains by 1908 worth nearly 10 times their original cost at par, while the average manufacturing stock appreciated only 300 percent with reinvestment in the same period. Compare U.S. Bureau of Corporations (1911, pp. 37–38) and Cowles (1938, pp. 170–71). Still, the evidence is incomplete and not altogether conclusive, especially since there were no artificial restrictions on entry into the snuff, tobacco, and petroleum industries. However a resolution of the conflict between "no successor competition" and "no monopoly in the first place" is not crucial in this instance, because neither inference is particularly favorable to the Justice Department.

[7] See Cox (1933, pp. 217–18), Nicholls (1951, p. 409), Stocking (1925, pp. 113–14), and Tennant (1950, p. 342). These and other studies of the tobacco and petroleum dissolutions are summarized in Whitney (1958, 1: 101–8, and 2: 14–22). Stigler (1956, p. 505) presents a brief but highly critical evaluation of industry studies.

[8] Dewey (1959, pp. 247–48). See also *Standard Oil Company of N.J. v. U.S.*, 221 U.S. 78 and *U.S. v. American Tobacco Co.*, 221 U.S. 185.

Likewise, studies of the price revisions accompanying stock splits, earnings declarations, and changes in accounting practices indicate that security prices adjust promptly and accurately to new information.[2] So far the research has concentrated on testing the validity of these models, but the accumulated evidence is now sufficiently strong that financial analysis can be profitably applied to other problems in economics. Thus, instead of determining that security prices react as expected to events whose consequences are largely known ex ante, their behavior can be used to identify the meaning and significance of the events themselves.

Industrial organization is a fertile field for such applications because the monthly character of stock prices makes them an effective substitute for annual measures of market performance which are often unavailable.[3] One topic that has generated new interest is the economic impact of trust-busting, and the experience with previous Sherman Act cases should provide a useful historical framework for evaluating current efforts (e.g., AT & T and IBM) to foster rivalry through dissolution. Consequently this study reexamines the competitive effects of the horizontal dissolutions carried out against Standard Oil, American Tobacco, and its subsidiary American Snuff on December 1, 1911.[4] The sample is small but surprisingly diverse, and the principal features of each decree are outlined in table 1.[5]

Security prices reveal the impact of dissolution quite dramatically. At a minimum, these data indicate that the performance of the snuff, tobacco, and petroleum industries was not significantly altered by the replacement of each trust with independent rivals. However, if one accepts the prevailing view that the three defendants were empirical examples of textbook monopolies, the results also imply that there was very little actual or anticipated competition between the firms in table 1 and no visible erosion of monopolistic rents.[6] This particular finding is identical to the conclusions of the dissolution literature, but it emerges without many of the ambiguities that characterize industry studies.[7] The use of security prices also clarifies what could be an inescapable dilemma of section 2 enforcement: balancing the consumer interest in vigorous rivalry with shareowner concern for the preservation of monopoly profits and capital values. Stockholder injury is an important consideration in dissolution proceedings, and the infrequent use of structural remedies is usually blamed on the reluctance of the courts to punish allegedly "innocent" investors.[8] In short, the results suggest that financial analysis

---

[2] See Ball (1972), Fama et al. (1969), and Jordan (1973).

[3] Previous uses of security prices in industrial organization include Ellert (1975), Lev and Mandelker (1972), Ornstein (1972), and Stigler (1965). Ellert's study is especially noteworthy because he used similar methods to analyze the relative price behavior of 1,466 common stocks issued by corporations involved in 1,019 antitrust cases between 1935 and 1971. Ellert also concludes (pp. 127, 129) that security prices are generally sensitive to new information with antitrust implications.

[4] *Standard Oil Company of N.J. v. U.S.*, 221 U.S. 1 (1911) and *U.S. v. American Tobacco Co.*, 221 U.S. 106 (1911). The complete texts of the dissolution decrees are presented in U.S. Courts (1918, pp. 136–44, 162–91). The snuff-tobacco decree is summarized in both the October 21, 1911 issue of the *Commercial and Financial Chronicle* (*CFC*), pp. 1122–24, and *Poor's Manual of Industrials*, 1912, pp. 1643–44, 1657–60.

[5] According to Whitney (1958, 2: 386–87), the Justice Department filed 17 Sherman Act suits against large manufacturing corporations between 1900 and 1920. The sample was restricted to the snuff, tobacco, and petroleum trusts primarily because of the expense of collecting substantial amounts of data from original sources. Studies of financial events in more recent periods typically avoid this problem with the use of the stock price tapes compiled by the Center for Research in Security Prices (CRSP) at the University of Chicago; this series begins in January 1926 as described in Fisher and Lorie (1964, pp. 2–3). Of the remaining cases in Whitney, only *U.S. v. E. I. du Pont de Nemours & Co.*, 188 Fed. 127 (D. Del. 1911) resulted in dissolution, and it was excluded due to insufficient stock price data. The other suits produced acquittals or minor consent decrees. Thus, the sample covers all but one of the early manufacturing dissolutions, including the landmark cases of *Standard Oil* and *American Tobacco*, but the empirical results are not necessarily descriptive of all antitrust suits filed during the period 1900–1920.

---

[6] Unfortunately, as is often the case in industrial organization, the empirical results will support other inferences besides "no competition between the successor companies." For example, it is conceivable that the histories of the snuff, tobacco, and petroleum trusts contained in U.S. Bureau of Corporations (1907 and 1909–15) overestimated both their market power and the extent of output restriction. If so, then the absence of negative effects from dissolution in the stock price data suggests that these industries were essentially competitive prior to 1912. Williamson and Andreano (1965) argue that such misconceptions are prevalent in discussions of Standard Oil. The available evidence for all three defendants is reviewed in Burns (1975, pp. 12–36), and some of it is consistent with the conventional view. Thus the defendants controlled substantial shares of most industry branches:

American Snuff (1900–1910)		American Tobacco (1900–1910)		Standard Oil (1899–1911)	
Snuff	91%	Cigarettes	85%	Kerosene	80%
		Little cigars	79	Gasoline	75
		Chewing	77	Waxes	59
		Fine-cut	74	Fuel oil	58
		Smoking	69	Lubricants	47
		Cigars	13		

U.S. Bureau of Corporations (1915, p. 2) and Williamson and Andreano (1965, p. 309). Likewise, the Bureau of Corporations reported that the purchase of American Tobacco common shares in 1890 would have generated income and capital gains by 1908 worth nearly 10 times their original cost at par, while the average manufacturing stock appreciated only 300 percent with reinvestment in the same period. Compare U.S. Bureau of Corporations (1911, pp. 37–38) and Cowles (1938, pp. 170–71). Still, the evidence is incomplete and not altogether conclusive, especially since there were no artificial restrictions on entry into the snuff, tobacco, and petroleum industries. However a resolution of the conflict between "no successor competition" and "no monopoly in the first place" is not crucial in this instance, because neither inference is particularly favorable to the Justice Department.

[7] See Cox (1933, pp. 217–18), Nicholls (1951, p. 409), Stocking (1925, pp. 113–14), and Tennant (1950, p. 342). These and other studies of the tobacco and petroleum dissolutions are summarized in Whitney (1958, 1: 101–8, and 2: 14–22). Stigler (1956, p. 505) presents a brief but highly critical evaluation of industry studies.

[8] Dewey (1959, pp. 247–48). See also *Standard Oil Company of N.J. v. U.S.*, 221 U.S. 78 and *U.S. v. American Tobacco Co.*, 221 U.S. 185.

TABLE 1

HIGHLIGHTS OF THE SNUFF, TOBACCO, AND PETROLEUM DISSOLUTIONS
(December 1, 1911)

Dissolution	Type	Number of Successor Companies	Major Successors	Geographic Division?[a]
Snuff	Partition	3	American Snuff George W. Helme Weyman-Bruton	Yes
Tobacco	Partition and divestiture	11	American Tobacco Liggett & Myers P. Lorillard R. J. Reynolds	No
Petroleum	Divestiture	34	Standard Oil (N.J.) Standard Oil (Cal.) Standard Oil (Ind.) Standard Oil (N.Y.) Standard Oil (Ohio)	Yes

SOURCES.—Cox (1933, p. 234), Stocking (1925, pp. 68–69), and Tennant (1950, pp. 60–61).

[a] A geographic division refers to the creation of separate marketing territories, each dominated by one successor company.

can be an important addition to the empirical research techniques of industrial organization.

The discussion is organized as follows. Section II briefly describes the advantages and limitations of a financial approach to trust-busting. A simple model of dissolution is developed in Section III. It provides various formulas for assembling the securities of each trust-successor group into portfolios that are comparable over time. Section IV describes the statistical tests of the decrees' competitive effects, Section V analyzes the results, and Section VI presents some additional conclusions.

II. The Case for Security Prices

Stock and bond prices have two major advantages over such conventional measures of resource allocation as output, product prices, and accounting rates of return. In the first place, they permit a direct comparison of trust and successor performance. This is the best way to test the impact of dissolution, but it is not feasible with traditional methods because the data no longer exist in sufficient quantity. The history of the trust supplies benchmarks of any restrictive behavior, and economic theory predicts that increased competition will be accompanied by lower product prices, greater outputs, and reduced earnings. *Ceteris paribus*, these changes imply correspondingly lower monthly stock prices. Conventional performance measures usually rely on annual data, but here there are as many as 12 observations per year. Consequently, the outcome of each suit can be identified without extending the analysis to more distant periods where extraneous developments are likely to obscure the results. The use of security prices also enhances the objectivity of the empirical tests, because observed capital values were determined by the buying and selling activities of shareowners with strong financial interests in correct judgments. Since there is no reason to believe that early-twentieth-century investors evaluated the decrees less accurately than modern economists, the present task is reduced to assembling the data properly and then selecting empirical techniques from corporate finance which will record the reactions of these traders to dissolution.

On the other hand, stock and bond prices are notoriously volatile. Time series return distributions have large variances, and current stochastic models are obviously incomplete.[9] As a result, the price behavior of trust and successor securities will reflect general market noise along with the influence of major developments in both the litigation and the reorganized industry. Fortunately, these random errors will tend to cancel out over time, and the magnitude of the problem can be reduced further by: (1) restricting the analysis of specific price changes to months where there is a strong a priori expectation of systematic effects; and (2) evaluating the significance of each test statistic against a distribution of equivalent values compiled for the securities of contemporary manufacturing corporations.

III. A Financial Model of Dissolution

Trust-busting always involves a division of production among independent centers of authority and a simultaneous rearrangement of ownership claims for each class of securities. The defendant relinquishes some of its property while shareholders typically receive pro-rata allotments of successor stocks. Both transfers occur regardless of the actual procedures used to break up a convicted monopolist. Two common devices are: *partition*, the financing of new firms to operate the segregated assets of the trust; and *divestiture*, the forced distribution (usually as dividends) of shares in various subsidiary corporations. The distinction between property and claims transfers is important because the outcome of dissolution depends exclusively on the reorganization of factories, products, and markets—and not on the disposition of securities.[10] Increased competition following the decree occurs when the successor companies have the capacity to sell close substitutes in contiguous territories but fail to maintain collusive agreements. These characteristics of the new industry are determined, at least initially, by the court's

[9] Fama (1965, pp. 98–99) and Fama et al. (1969, p. 7).

[10] Qualifications may be necessary when the exchange of shares drastically alters the distribution of stockholder control over the properties in litigation. However, the courts did not seriously dilute the voting power of the principal trust owners in the reorganized snuff, tobacco, and petroleum industries; see Burns (1975, pp. 17, 24–25, 31, 181–83).

division of assets, and they are unaffected by the capital structures of the firms involved in the litigation or the corresponding terms of share exchange.

Comparable portfolios of trust and successor securities can be assembled systematically from the valuation model formulated by Schall (1972).[11] In a perfect capital market with zero transactions costs and taxes, homogeneous expectations, and the opportunity to borrow and lend at the same interest rate, his Proposition I (p. 13) ensures that the market value of a sum of income streams is equal to the sum of their market values.[12] Thus, if X_i refers to the expected earnings of the ith asset and V_i is its corresponding market value, then $X = \sum_i X_i$ implies

$$V = \sum_i V_i \tag{1}$$

in equilibrium. Schall also demonstrates (pp. 13–18) that V is independent of both the covariance between separate components of the firm's income and the relative risks of the activities which generate them, because traders can replicate desired patterns of corporate diversification in their personal accounts.

These results are easily applied to dissolution. Suppose there are N successor companies and let the subscript $i = 0, 1, 2, \ldots, N$ denote individual firms, where $i = 0, 1$ for the trust and its direct successor, respectively. The financial changes accompanying *partition* do not alter group earnings, so

$$X_0 = \sum_{i=1}^{N} X_i.$$

Therefore the creation and cancellation of securities preserves the intrinsic total market value of the properties in litigation such that

$$V_0 = \sum_{i=1}^{N} V_i.$$

Now define $E_i \equiv n_i p_i$ = market value of equity and $D_i \equiv n_i^* p_i^*$ = market value of debt, where p_i = price per share of common stock, n_i = number of common shares outstanding, p_i^* = price per bond, and n_i^* = number of bonds outstanding. Then $V_i \equiv E_i + D_i \equiv n_i p_i + n_i^* p_i^*$ for all i, and rearranging

$$V_0 = \sum_{i=1}^{N} V_i$$

---

[11] Identical results can be obtained with somewhat greater effort by using either the risk class theory of valuation or the CAPM; see Modigliani and Miller (1958), Hamada (1969), Schall (1972, p. 21, n. 17), and Burns (1975, pp. 41–55).

[12] The one item in this list which can be verified empirically is the assumption of zero taxes. According to Pechman (1966, pp. 243, 245), the range of personal and corporate tax rates was only 0–15 percent and 0–2 percent, respectively, during the 1904–16 sample period.

gives

$$p_0 + \frac{n_0^*}{n_0} p_0^* = \sum_{i=1}^{N} \left[\frac{n_i}{n_0} p_i + \frac{n_i^*}{n_0} p_i^* \right]. \tag{2}$$

For pure *divestiture* write $X_i = (X_i - r_i n_i^*) + r_i n_i^*$ where r_i = bond coupon rate (in dollars), and let $0 \leq \theta_i, \theta_i^* \leq 1$ be the trust's prior fractional ownership of the respective equity and debt issued by the newly liberated subsidiaries, firms 2, 3, . . . , N. Then

$$X_0 = X_1 + \sum_{i=2}^{N} \theta_i (X_i - r_i n_i^*) + \sum_{i=2}^{N} \theta_i^* r_i n_i^*$$

and (1) becomes

$$V_0 = V_1 + \sum_{i=2}^{N} (\theta_i E_i + \theta_i^* D_i)$$

or

$$p_0 + \frac{n_0^*}{n_0} p_0^* = p_1 + \frac{n_0^*}{n_0} p_1^* + \sum_{i=2}^{N} \left[\theta_i \frac{n_i}{n_0} p_i + \theta_i^* \frac{n_i^*}{n_0} p_i^* \right] \tag{3}$$

since $n_1 = n_0$ and $n_1^* = n_0^*$.

Equations (2)–(3) provide general formulas to make the transition between trust and successor securities on December 1, 1911. These portfolios duplicate the total market value of the defendants' assets, and they take into account all required capital changes. The result in each case is a virtually complete set of comparable monthly data which surrounds the decree and reflects investor expectations of dissolution. The equations also can be combined for the joint imposition of partition and divestiture, but they must be adjusted for stock splits, debt retirement, etc., as discussed in Burns (1975, pp. 184–222). The only noteworthy adaptation is the elimination of American Tobacco's (AT) prior ownership of American Snuff (AS). Rewriting (3) as

$$\left[p_0 + \frac{n_0^*}{n_0} p_0^* \right]_{AT-AS} = \left[p_0 + \frac{n_0^*}{n_0} p_0^* \right]_{AT} - \left[\frac{1}{n_0} \right]_{AT} [\theta n_0 p_0 + \theta^* n_0^* p_0^*]_{AS} \tag{4}$$

gives the computed value of the tobacco portfolio without American Snuff. Equation (4) makes the empirical tests of snuff and tobacco securities independent of one another.

IV. Tests of Competitive Effects

Dissolution is one of several factors which influence the value of the portfolios in (2)–(3). A systematic determinant of monthly price changes is the level of marketwide activity, and a popular form of the stochastic relationship between individual and market returns is[13]

$$R_{jt} = \alpha_j + \beta_j R_{mt} + \varepsilon_{jt} \tag{5}$$

where R_{jt} = the return during month t on the jth trust-successor portfolio as given by (2) or (3); R_{mt} = the corresponding rate of return for all stocks; $\beta_j = \text{cov}(R_j, R_m)/\sigma^2(R_m)$; and ε_{jt} = random error. Equation (5) is consistent with the CAPM under the assumption that stock returns have a stationary multivariate normal distribution.[14] The slope coefficient β_j measures relative risk in the CAPM, and these values must sum to unity when combined using the same weights that produce R_{mt} from the returns on individual securities. This market model underlies the following tests of competitive effects based on developments in the litigation.

The portfolio values in (2)–(3) will be revised upward or downward relative to the market average as new information about the firms becomes available to investors. Dissolution proceedings are accompanied by four events which generate the kind of new information that should have an impact on observed prices. They are: the filing of a Sherman Act complaint (in month t_1), the Supreme Court decision (t_2), publication of the dissolution plan (t_3), and the initial dividend declarations of the successor companies (t_4). The significance of these events is fairly obvious. The beginning of the antitrust suit and the final court decision ordering dissolution create expectations that future profits will be reduced by intensified competition. Publication of the dissolution plan gives investors their first look at the new structure of the industry. This information is often sufficiently comprehensive to support inferences about probable conduct and performance effects.[15] Finally, the initial dividend announcements reveal, however vaguely, both the actual experience of the successor companies as well as their forecasts of long-run earnings.[16]

Least-squares estimates of (5) over the months preceding the complaint

$$R_{jt} = \hat{\alpha}_j + \hat{\beta}_j R_{mt} + u_{jt} \qquad t = 1, 2, \ldots, t_1 - 1 \tag{6}$$

provide the desired benchmark of trust performance, as reflected in the prices of its securities, because (6) is unaffected by the litigation.[17] Therefore, the impact of those four events can be isolated from both market trends and the specific risk of the jth trust-successor portfolio by extrapolating (6) into subsequent months and computing the residuals

---

[13] See Fama et al. (1969, p. 4), Jordan (1973, p. 611), and Lev and Mandelker (1972, p. 96).

[14] The links between (5) and the CAPM are discussed in Fama (1973).

[15] For example, the summary of the snuff-tobacco dissolution plan published in the *CFC* (October 21, 1911), pp. 1122–24, identifies all 14 successor companies, the percentage of each new firm owned by the principal trust stockholders, the expected market shares of the major successors in 1912, the product brands awarded to them, and the location of their factories.

[16] See Fama et al. (1969, pp. 2–3) and the references cited therein.

[17] There is no evidence that news of the petroleum and snuff-tobacco complaints was leaked to the investment community prior to month t_1 (November 1906 and July 1907, respectively). The weekly "Intelligence Reports" of the *CFC* do not contain any information during the preceding 3 years which would have suggested that these prosecutions were imminent.

$$\hat{u}_{jt} = R_{jt} - (\hat{\alpha}_j + \hat{\beta}_j R_{mt}) \qquad t = t_1, t_1 + 1, \ldots, T \tag{7}$$

where $T > t_4$.[18] These values correspond to the market-adjusted capital gains and losses arising from new information.[19] Previous discussion implies that

$$E[\hat{u}_{jt_1}], E[\hat{u}_{jt_2}] < 0 \tag{8}$$

and the alternative investor evaluations of the dissolution plan and successor dividend announcements are

$$E[\hat{u}_{jt_3}], E[\hat{u}_{jt_4}] \begin{array}{l} < 0\text{: more severe than anticipated} \\ = 0\text{: bad as expected} \\ > 0\text{: better than expected.} \end{array} \tag{9}$$

Since the four events may produce offsetting relative price changes, an appropriate test of the overall impact of trust-busting is

$$E[\hat{u}_{jt_1} + \hat{u}_{jt_2} + \hat{u}_{jt_3} + \hat{u}_{jt_4}] \begin{array}{l} \geq 0\text{: benign} \\ < 0\text{: competitive effects.} \end{array} \tag{10}$$

Non-negative residual sums indicate that capital gains during t_3 and t_4 tended to reestablish the predissolution relationship in (6) between trust and market returns. The results also can be displayed graphically for the period immediately surrounding each decree by calculating the cumulative residuals

$$\hat{U}_{j\tau} = \sum_{t = \text{Jan. } 1911}^{\tau = \text{Dec. } 1912} \hat{u}_{jt}. \tag{11}$$

This interval contains t_2, t_3, and t_4, so major shifts in the trend of $\hat{U}_{j\tau}$ should coincide with the Supreme Court decision, publication of the dissolution plan, and the initial dividend declarations.

The interpretation of (8)–(10) requires evidence that the snuff, tobacco, and petroleum residuals differ substantially from the ordinary experience of contemporary manufacturing corporations, because the litigation and its aftermath are not the only sources of new information which determine the relative price changes of trust-successor securities. The $\hat{u}_{jt}$'s in (7) are influenced by random events along with variables that were omitted from the market model, and these errors must be accounted

---

[18] This approach was first developed by Fama et al. (1969, pp. 7–9).

[19] Let $R_{jt} = \ln (A_{jt}/A_{j,t-1})$ where A_{jt} is the value of the portfolio in (2) or (3) with reinvestment of all cash receipts in additional shares. Then (7) can be written as

$$\hat{u}_{jt} = \ln A_{jt} - (\ln A_{j,t-1} + \hat{\alpha}_j + \hat{\beta}_j R_{mt}) \qquad t = t_1, t_1 + 1, \ldots, T$$

which is the natural logarithm of the market-adjusted gain or loss brought on by new information in month t.

for in the residual tests. With this in mind, the calculations of Sections III–IV were replicated for the shares of 28 other industrial firms which satisfied the following criteria: (1) the corporation issued both debt and equity; (2) these securities were listed continuously on either the outside market or the New York Stock Exchange (NYSE) between March 1904 and December 1916; and (3) the firm was not subjected to bankruptcy or a major antitrust prosecution. Extrapolations of

$$R_{ct} = \hat{\alpha}_c + \hat{\beta}_c R_{mt} + u_{ct} \qquad \begin{matrix} t = 1, 2, \ldots, t_1 - 1 \\ c = 1, 2, \ldots, 28 \end{matrix}, \tag{12}$$

which is the control group equivalent of (6), generate approximately normal distributions of estimated residuals during t_1, t_2, t_3, and t_4. Consequently, the statistical significance of the $\hat{u}_{jt}$'s in (8)–(10) can be evaluated against the mean and standard deviation of the corresponding distributions assembled from the securities of the 28 other companies. A full description of the control group is presented in the Appendix, along with a discussion of the cross-section independence and normality of the sample.

Two potential limitations of the empirical tests should be mentioned briefly. First, the behavior of $\hat{u}_{jt_1}$ and $\hat{u}_{jt_2}$ may reflect the increased uncertainty of investors regarding the future of each trust as well as expectations of reduced earnings—since both the complaint and the decision create important issues which are temporarily left unresolved. This argument implies that capital gains could accompany publication of the dissolution plan and the successor dividend announcements solely because the events dispel uncertainty. Interpretations of (9) are clouded by the possible compounding of these effects, although (10) remains satisfactory because the residual sums incorporate all four events. Previous studies have not addressed this problem, and no empirical methods are available to cope with it. However, widespread panic during t_1 and t_2 seems unlikely when traders have a vested interest in the dispassionate analysis of new information. The other difficulty concerns the extent to which investors anticipated the major developments associated with dissolution, so that the corresponding price revisions actually occurred before t_1, t_2, t_3, and t_4, respectively. The tests in (8)–(10) are inappropriate if the complaint, decision, dissolution plan, and dividend announcements were correctly forecast *ex ante*. This outcome seems remote in general, because there is no evidence of leakage from the litigation, and, in contrast to studies of purely financial events, there does not appear to be an alternative source of new information that would have consistently supported prior expectations. The issue can be largely resolved by examining the cumulative residuals in (11), since their trend should be relatively constant for $\tau \neq t_2$, t_3, and t_4 when anticipations are negligible.

V. Empirical Results

All rates of return were computed as the natural logarithm of the ratio of portfolio values with reinvestment during successive months:

$$R_{jt} = 100 \ln \frac{A_{jt}}{A_{j,t-1}} \quad \text{and} \quad R_{mt} = 100 \ln \frac{A_{mt}}{A_{m,t-1}}$$

where A_{jt} = the value of the jth trust-successor portfolio (from [2] or [3]) in month t with reinvestment of all cash receipts in additional shares, and A_{mt} = the corresponding value of the Cowles Index of all common stocks,[20] and likewise for R_{ct}. It was necessary to simulate the reinvestment of cash distributions because A_{mt} was constructed on this basis, as discussed in Cowles (1938, pp. 3–25). Ordinarily the procedure is a simple application of compound interest.[21]

The portfolio accumulations begin in March 1904 and usually end in December 1916, although the interval of primary interest is January 1911–December 1912. All three defendants were well established prior to 1904, and there are at least 31 return observations for trust securities prior to the filing of a Sherman Act complaint.[22] In order to synchronize the measurement of R_{jt} and R_{ct} with R_{mt}, the prices of shares listed on the NYSE were computed as the average, adjusted for dividends and other payments, of the monthly range published in the *Bank and Quotation*

---

[20] Cowles (1938, table C-1, pp. 168–69). Lorie and Hamilton (1973, p. 30) describe it as the "best" of the early stock market indices, and these are surveyed in Cowles (1938, pp. 437–48). The components of the Cowles Index are weighted by their market values which increases the influence of the returns for large corporations relative to an index of equally weighted securities. Consequently the index of all common stocks (industrials, railroads, and public utilities) was selected to minimize the effects of an individual R_{jt} (or R_{ct}) on the value-weighted R_{mt}.

[21] Suppose dissolution takes the form of partition and let

$$P_{jt} = \left[p_{0t} + \frac{n^*_{0t}}{n_0} p^*_{0t}\right], \sum_{i=1}^{N} \left[\frac{n_{it}}{n_0} p_{it} + \frac{n^*_{it}}{n_0} p^*_{it}\right]$$

= the price per share of the jth trust-successor portfolio from (2)

S_{jt} = the number of portfolio shares owned in month t

$$C_{jt} = \left[c_{0t} + \frac{n^*_{0,t-1}}{n_0} c^*_{0t}\right], \sum_{i=1}^{N} \left[\frac{n_{i,t-1}}{n_0} c_{it} + \frac{n^*_{i,t-1}}{n_0} c^*_{it}\right]$$

= the cash receipt per share of the jth portfolio

where c_{it} and c^*_{it} are the respective distributions per common stocks and debt securities outstanding during the previous month. Then $S_{jt} = S_{j,t-1}(1 + C_{jt}/P_{jt})$ and $A_{jt} = S_{jt}P_{jt}$ where $S_{j0} = 1.0$. The adjustment of these formulas for stock splits, debt retirement, spin-offs, and subscriptions to new issues is described in Burns (1975, pp. 223–31).

[22] Detailed histories of Standard Oil and American Tobacco–American Snuff are presented in U.S. Bureau of Corporations (1907 and 1909, respectively).

TABLE 2

REGRESSION RESULTS: TRUST PORTFOLIOS OF TOTAL MARKET VALUE

Company	Number of Observations	Period	$\hat{\alpha}_j$	$\hat{\beta}_j$	R^2	F	d
American Snuff	39	April 1904–June 1907	1.104 (1.54)	0.583(2.80)	.175	7.85	2.45
American Tobacco	39	April 1904–June 1907	0.622 (1.19)	0.596(3.96)	.297	15.65	1.89
Standard Oil	31	April 1904–October 1906	−0.269(−0.548)	0.328(2.20)	.143	4.85	1.88

NOTE.—The figures in parentheses are t-values.

TABLE 3

SUMMARY OF REGRESSION RESULTS:
TOTAL MARKET VALUE PORTFOLIOS OF 28 OTHER COMPANIES

PERIOD AND STATISTIC	RANGE Low	RANGE High	MEAN	MEDIAN	STANDARD DEVIATION
April 1904–June 1907:					
$\hat{\alpha}_c$	−2.281	1.866	0.404	0.591	0.989
$\hat{\beta}_c$	0.189	1.822	0.869	0.835	0.454
R^2	.039	.688	.379	.396	.165
F	1.51	81.77	27.28	24.21	19.44
d	1.05	2.29	1.66	1.67	0.322
April 1904–October 1906:					
$\hat{\alpha}_c$	−2.096	2.088	0.424	0.562	1.122
$\hat{\beta}_c$	0.147	2.418	0.872	0.763	0.553
R^2	.015	.609	.283	.266	.161
F	0.45	45.10	13.77	11.46	11.24
d	1.00	2.35	1.65	1.68	0.332

Supplement to the *Commercial and Financial Chronicle.*[23] Certain other stocks were listed only on the outside market and were not traded regularly. The prices of these securities were computed as the adjusted average of the bid and ask quotations during the first and fourth week of each month. Both methods of compiling observed prices typically create a 30-day lag in the full response of portfolio values to particular events, and the residuals in (8)–(10) must be accumulated accordingly. No prices exist for August–December 1914 (NYSE closed), and none are available for the successor companies during December 1911–February 1912 (exchange of shares). *Moody's* and *Poor's* industrial manuals supplied the remaining data on cash distributions, shares outstanding, and trust stock holdings.

Table 2 presents the least-squares estimates of (6) for the trust portfolios, and table 3 summarizes the corresponding results from (12) for the control group. Every $\hat{\beta}$ has the expected positive sign and 86 percent (51/59) of them are statistically different from zero at the 95 percent confidence level.[24] These results are especially robust in light of the fact that the total market value portfolios generally contain large amounts of debt, which tends to reduce estimates of specific risk.[25] There is no indication of heteroscedasticity in the scatter diagrams of R_{jt} and R_{ct} on R_{mt}, and the Durbin-Watson statistics show little evidence of autocorrelation. In short, tables 2–3 support the view that market trends exert a major influence on the returns of individual securities.[26]

---

[23] The use of mid-range sale prices is essential because the Cowles Index was derived from them, as were the majority of stock market averages which include the sample period. In addition, none of the alternative indices are comparable to A_{mt} for coverage of NYSE securities; see Cowles (1938, pp. 5, 17, 437–48). However, this convention raises a problem in that the research on security returns has been carried out almost exclusively with daily or end-of-month data, so the recognized properties of market model regressions may not be strictly applicable to the coefficient estimates presented in the text. But potential discrepancies are unavoidable anyway, because previous studies also focused solely on the behavior of common stock returns in more recent periods than 1904–16; see Fama (1965, p. 45), Fama and MacBeth (1973, p. 614), Fama et al. (1969, p. 3), Blattberg and Gonedes (1974, p. 261), and Martin and Klemkosky (1975, pp. 82–83).

Table 4 presents the reactions of investors to the new information generated by the litigation, and the residual patterns are fairly similar in all three cases. The filing of complaints against Standard Oil and American Tobacco in November 1906 and July 1907, respectively, produced negative and statistically significant departures from previous relationships between trust and market returns—as traders began anticipating dissolution and revising prices downward. This event reduced the value of snuff, tobacco, and petroleum securities by 9–16 percent.[27] Further declines followed the Supreme Court decision in *American Tobacco* on May 29, 1911, when the government's victory reinforced earlier fears in the financial community. Allowing for market trends, the aggregate value of snuff securities fell more than 20 percent, while the prices of tobacco shares declined by almost 8 percent. Both reductions are significant at the 95 percent level. In contrast, the petroleum residuals are quite small during May and June 1911, probably because the Supreme Court opinion contained a relatively benign dissolution plan.[28]

The declining prices of trust securities were subsequently reversed by news of the reorganization in each industry. Publication of the snuff and tobacco dissolution plans in October 1911 pushed the prices of American Snuff shares upward by almost 10 percent relative to the market average. An additional gain of 15 percent followed the first announcement of

---

[24] Previous studies indicate that distributions of daily stock returns exhibit longer tails than the normal, although conventional statistical testing with an assumption of approximate normality is still appropriate; see Fama (1965, pp. 42–55), Fama and MacBeth (1973, pp. 619–24), and Fama et al. (1969, pp. 6–7). Blattberg and Gonedes (1974) report that the Student t distribution explains these thick tails, and this result implies that monthly return distributions will be approximately normal.

[25] When (12) was estimated over April 1904–June 1907 for the common stocks in the control group, the mean β_c increased more than 60 percent to 1.452 with a standard deviation of .556.

[26] Similar results were obtained from a sample of 622 common stocks by Fama et al. (1969, pp. 3–6). They performed 940 estimates of (5) and found a mean β_j equal to .894 with a standard deviation of .305. The mean R^2 was .399. They also reported no evidence of heteroscedasticity, nonlinearity, or serial dependence.

[27] Market-adjusted price changes reported as percentages in the text are equivalent to the corresponding residual entries in table 4, because logarithmic and percent transformations of portfolio values yield approximately the same results for changes of less than 15 percent; see Blattberg and Gonedes (1974, pp. 252–53).

[28] *Standard Oil Company of N.J. v. U.S.*, 221 U.S. 77–82 (1911).

TABLE 4

ESTIMATED RESIDUALS: TRUST PORTFOLIOS OF TOTAL MARKET VALUE

TRUST AND EVENT	MONTH(S)	RESIDUAL(S)[a]	CONTROL GROUP RESIDUAL DISTRIBUTION Mean	CONTROL GROUP RESIDUAL DISTRIBUTION Standard Deviation[b]
Snuff:				
Complaint	July 1907	−10.840*	−0.945	2.558
	July, August 1907	−11.393*	−1.435	5.327
Decision	June 1911	−20.854*	−0.344	3.756
Dissolution plan....	October 1911	2.493	−0.867	3.699
	October, November 1911	9.790‡	−2.398	5.031
Preferred dividends	March 1912	−3.296	3.276	7.656
	March, April 1912	−2.700	5.071	9.348
Common dividends	June 1912	6.950‡	−0.940	3.334
	June, July 1912	15.051‡	−1.246	5.870
All	July, August 1907 June 1911 October, November 1911 March, April 1912 June, July 1912	−10.106	−0.353	18.988
Tobacco:				
Complaint	July 1907	−6.328*	−0.945	2.558
	July, August 1907	−16.096*	−1.435	5.327
Decision	June 1911	−7.809*	−0.344	3.756
Dissolution plan....	October 1911	−0.329	−0.867	3.699
	October, November 1911	4.373	−2.398	5.031
Preferred dividends	March 1912	12.612	3.276	7.656
	March, April 1912	14.482	5.071	9.348
Common dividends	May 1912	1.308	2.049	4.428
	May, June 1912	7.999§	1.109	4.605
	May, June, July 1912	9.706§	0.803	6.011
All	July, August 1907 June 1911 October, November 1911 March, April 1912 May, June, July 1912	4.656	1.705	21.071
Petroleum:				
Complaint	November 1906	−6.473*	−0.727	2.707
	November, December 1906	−9.021†	−0.845	5.364
Decision and dissolution plan..	May 1911	−0.218	−0.410	2.746
	May, June 1911	−1.871	−0.779	5.447
Common dividends	March 1912	35.229‡	3.247	7.706
	March, April 1912	37.427‡	5.013	9.284
All	November, December 1906 May, June 1911 March, April 1912	26.535§	3.389	14.122

---

[a] The residuals are accumulated over successive months because the impact of a particular event on trust and successor security prices ordinarily is carried over into the following month. Thus the residual entry for July, August 1907 is $(\hat{u}_{jt_1} + \hat{u}_{j,t_1+1})$, and it is $(\hat{u}_{jt_3} + \hat{u}_{j,t_3+1})$ for October, November 1911, etc.

[b] Computed as $[(1/27)\,\Sigma_{c=1}^{28}\,(\hat{u}_{ct} - \bar{u}_t)^2]^{1/2}$ for the corresponding month(s). An adjustment for the covariance within the control group residual distributions is described in the Appendix

* Statistically significant at the 95% confidence level on a one-tailed test

† Statistically significant at the 90% confidence level on a one-tailed test

‡ Statistically significant at the 95% confidence level on a two-tailed test.

§ Statistically significant at the 85% confidence level on a two-tailed test.

dividends on the common stock by all three successor companies during June 1912. These advances are statistically significant at the 95 percent level of confidence. The absence of price data for December 1911–February 1912 makes the tobacco residuals more difficult to interpret, and the problem is compounded by the fact that the major successors did not announce their initial common dividends in the same month.[29] However, the 9 percent cumulative advance in the prices of tobacco securities during the spring of 1912, resulting from dividend payments by American, Lorillard, and Reynolds, is significant at the 85 percent level. But by far the greatest increase in relative prices coincided with the dividend announcements of the Standard Oil successors. Fifteen of these companies, representing 75 percent of the petroleum portfolio in March 1912, declared at least one common dividend between November 1911 and April 1, 1912. The result was a market-adjusted increase of more than one-third in the value of Standard Oil stocks during this period.

The offsetting signs of the relative price changes that surround October 1, 1911 produce residual sums from (10) which are either positive or insufficiently negative. Thus, table 4 indicates that investors evaluated the snuff, tobacco, and petroleum dissolutions as "better than expected" and ultimately "benign" in their competitive effects.

An intuitive explanation for these residual patterns is that previously adverse forecasts of group profits were made obsolete by the new information disclosed in the dissolution plan and/or the dividend declarations of the successor companies.[30] Consequently, prices throughout the summer

---

[29] As noted previously in this section, the prices of snuff, tobacco, and petroleum securities do not exist for December 1911–February 1912, due to the exchange of shares accompanying each dissolution. These data are available for the 28 other companies, but the prices must be deleted so that their residuals correspond to those of the trusts and successors. Consequently, (7) treats November 1911 and March 1912 as adjacent months when calculating all of the estimated residuals. This procedure increases the variance within the control group distributions for March 1912 relative to that in any other single month, as can be seen in table 4. There is no problem in evaluating the snuff and petroleum residuals for March and April 1912, because the former are quite small, while the latter are sufficiently large to overcome the increased dispersion of the control group distributions. But the corresponding tobacco residuals also are substantial, and table 4 suggests that they might have been statistically significant at an acceptable level under more favorable circumstances. These values presumably reflect the information content of the successors' preferred dividend announcements, as well as the initial declaration of common dividends by R. J. Reynolds in December 1911. The major tobacco successors announced the following dividends payable to common stockholders (per share owned) during their first year of independent operation: Reynolds—December 1911 ($3), June and December 1912 ($5 each); American—May, August, and November 1912 ($2.50 each); Lorillard—June and September 1912 ($2.50 each), December 1912 ($5); and Liggett & Myers—November 1912 ($3); see *Poor's Manual of Industrials*, 1913, pp. 456, 996, and *CFC* (March 20, 1912), p. 921, (June 15, 1912), p. 1607, (November 2, 1912), p. 1175, and (December 14, 1912), p. 1588.

[30] Dividend announcements are apparently not proxies for alternative sources of new information. The weekly "Intelligence Reports" of the *CFC* do not contain interim earnings statements for any successor company during the first 7 months of 1912.

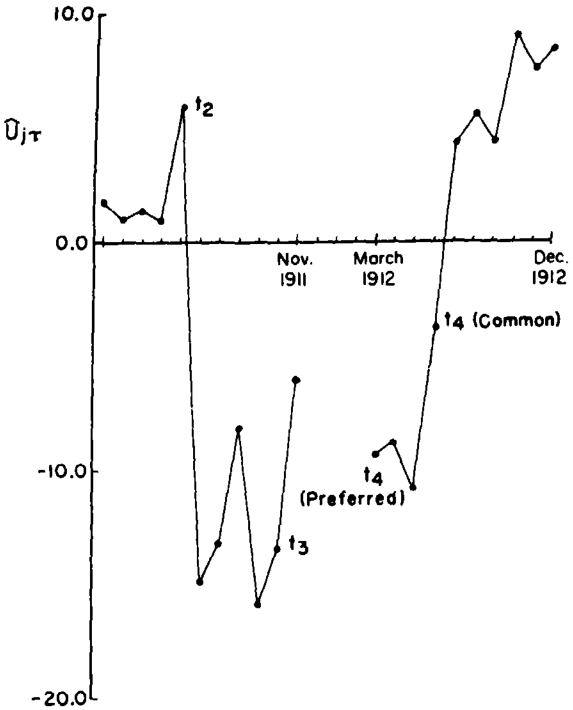


Fig. 1.— Cumulative residuals from the snuff portfolio of total market value, January 1911–December 1912.

of 1911 were below levels consistent with reorganizations that were no longer expected to threaten discounted earnings per share. As investors sought to acquire these securities at bargain prices, they bid up their value until the yield from successor stocks after the initial dividend announcements was approximately equal to the yield on the corresponding trust shares immediately following the Supreme Court decisions.

The cumulative residuals from (11) are presented in figures 1–3 for the snuff, tobacco, and petroleum companies, respectively. The break in each trend marks the transition between a trust and its successors. The general picture is one of capital gains during the first half of 1912 eliminating previous losses from the complaint and decision—and sending each series toward a new plateau which reflects record prices by the end of the year. This result is quite dramatic in all three cases, but otherwise there are some individual differences. Figure 3 shows the negligible impact of the *Standard Oil* decision as $\hat{U}_{jt}$ is relatively constant throughout 1911. Likewise, a comparison of figures 1–2 indicates that the tobacco

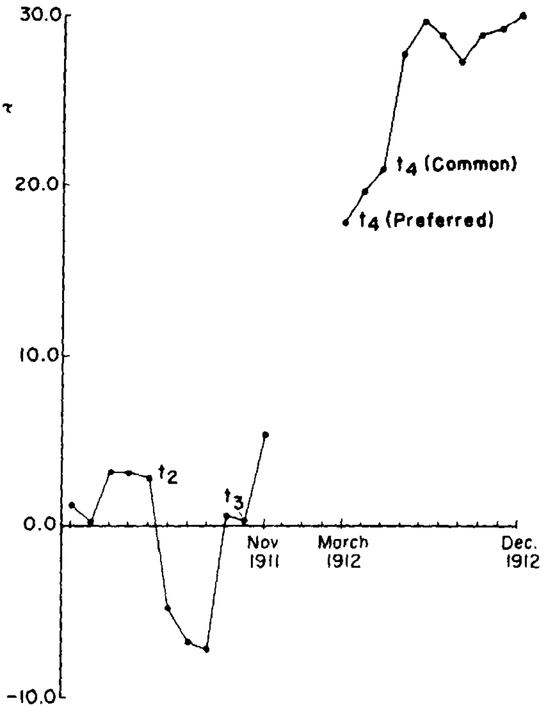


FIG. 2.—Cumulative residuals from the tobacco portfolio of total market value, January 1911–December 1912.

opinion imposed losses on American Snuff stockholders that were more immediate and proportionately greater than the price declines of American Tobacco securities. They also contain two possible instances of the modest anticipation of subsequent events: the rising trend during the month prior to t_4 (May 1912) in figure 1, and the capital gain 2 months before t_3 (August 1912) in figure 2. However, all three sets of cumulative residuals suggest that such advances were not uncommon in 1911–12 because of random errors in the market model.

VI. Conclusions

Subject to the qualifications noted in both Section IV and the Appendix, the evidence indicates that investors were pleasantly surprised by the dissolutions of American Snuff, American Tobacco, and Standard Oil. Empirical tests of the stock price revisions accompanying important developments in the litigation imply that each decree merely legitimized prevailing rates of output restriction and excess profits. Of course, in later

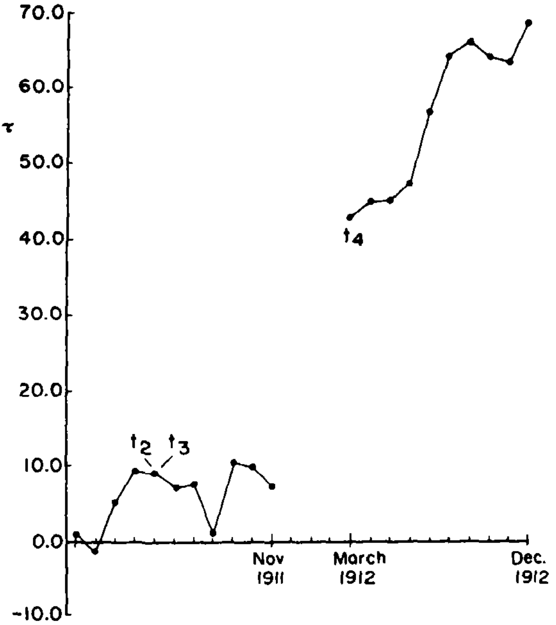


FIG. 3.—Cumulative residuals from the petroleum portfolio of total market value, January 1911–December 1912.

years competitive forces might have intensified in these industries due in part to the dismantling of the snuff, tobacco, and petroleum trusts. But astute traders would have anticipated this contingency in 1911–12, and evidently they decided it was remote enough not to affect the value of discounted earnings per share.

The correspondence between these findings and the conclusions of the dissolution literature means that little can be added to previous recommendations for more effective trust-busting. Certainly the courts should try to avoid geographic divisions and small numbers of successor companies, since these alterations do not produce visible changes in resource allocation. It also seems clear that judges have experienced undue and largely self-inflicted agony over the fate of "innocent" shareowners, because dissolution, as historically practiced, does not reduce capital values.

On an analytical level, security prices appear quite sensitive to new information with antitrust implications. This characteristic, along with their widespread availability and the development of empirical models in finance, makes them a potentially useful data source for industrial organization and perhaps other fields as well. Among further applications

TABLE A1

Twenty-eight Manufacturing Corporations

Company (Abbreviation)	Primary Outputs
American Agricultural Chemical (AAC)	Fertilizer, glue, and gelatine
American Beet Sugar (ABS)	Sugar
American Car & Foundry (ACF)	Freight and passenger cars
American Cotton Oil (ACO)	Cotton seed products
American Hide & Leather (AHL)	Meat and leather
American Linseed (ALIN)	Linseed oil and varnish
American Locomotive (ALOC)	Locomotives
American Malting (AM)	Beer
American Smelting & Refining (ASR)	Metal products
American Type Founders (ATF)	Printing machinery and type
American Woolen (AW)	Wool products
American Writing Paper (AWP)	Stationery and envelopes
Borden's Condensed Milk (BCM)	Evaporated milk and candy
Crucible Steel (CS)	Steel products
Distillers Securities (DS)	Liquor
General Chemical (GC)	Chemicals
International Paper (IP)	Lumber, wood pulp, and newsprint
National Biscuit (NB)	Biscuits and crackers
National Enameling & Stamping (NES)	Metal products
National Lead (NL)	Lead products and vegetable oils
Otis Elevator (OE)	Elevators and lifts
Pressed Steel Car (PSC)	Freight cars
Railway Steel Spring (RSS)	Steel springs and related products
Republic Iron & Steel (RIS)	Steel products
Union Bag & Paper (UBP)	Pulp, paper, and paper bags
U.S. Cast Iron Pipe & Foundry (USCIP)	Iron pipe and steel products
U.S. Rubber (USR)	Rubber products
Virginia-Carolina Chemical (VCC)	Chemicals and fertilizer

of stock prices, one which may be especially promising is a study of the economic impact of prosecutions under section 1 of the Sherman Act.[31] There are a large number of these cases, and it is conceivable that restraints against collusive agreements have often been as innocuous as the remedies imposed for illegal monopolization.

Appendix

The 28 companies in the control group are listed in table A-1. The sample contains every manufacturing corporation which satisfied the selection criteria described in the text.[32] Table A1 includes some of the largest manufacturing firms of the early twentieth century, along with other companies that dominated their respective industries.[33] Thus some of these corporations were comparable to

---

[31] The study of section 1 enforcement by Ellert (1975, pp. 89–101) represents an excellent start in this direction.

[32] Since the stock prices are not on the CRSP tapes, the selection was restricted to levered firms in order to reduce the expense of collecting data for both the present study of total market value and future tests of individual debt and equity shares. All but one of the debt securities in the control group are preferred stocks.

[33] Compare Table A1 with the lists in Collins and Preston (1961, pp. 1005–11) and Scherer (1970, pp. 48, 105–6).

TABLE A2

SUMMARY OF RESIDUAL CORRELATION COEFFICIENTS FOR TWENTY-EIGHT OTHER COMPANIES

NUMBER OF FIRMS	SELECTION PROCESS	NUMBER OF CORRELATIONS	PERIOD	RANGE		MEAN	MEDIAN	STANDARD DEVIATION
				Low	High			
13	Competitors	11	I	−.058	.306	.160	.186	.122
			II	−.071	.313	.153	.147	.141
7....	Random	21	I	−.230	.292	.097	.124	.165
			II	−.127	.465	.104	.118	.136

NOTE.—Period I = November 1906–December 1916, period II = July 1907–December 1916

American Snuff, American Tobacco, and Standard Oil in terms of either absolute or relative size. Complete sets of monthly portfolio values are available on request.

Table A-1 covers a fairly broad range of manufacturing activities, but certain industries are more heavily represented than others: there are four corporations producing railway equipment (ACF, ALOC, PSC, and RSS); three in chemicals (AAC, GC, and VCC); three in steel products (CS, RIS, and USCIP); and two each in fabricated metals (ASR and NES), linseed oil (ALIN and NL), and paper products (IP and UBP). This listing may not be exhaustive, and the overrepresentation of these industries raises questions about the degree of cross-section independence within the control group distributions of (8)–(10). In particular, substantial correlations between $\hat{u}_{c_1t}$ and $\hat{u}_{c_2t}$, $c_1 \neq c_2$, tend to increase the variance of those distributions and thereby reduce the confidence levels of the results presented in table 4.

Since it is impractical to compute standard deviations that account for all 378 covariances among the 28 firms, the following procedures were used to quantify the extent of cross-section dependence and lessen its influence on the statistical tests. First the $\hat{u}_{ct}$'s generated by (12) were correlated for the 15 pairs of potential rivals during each of the periods November 1906–December 1916 (period I) and July 1907–December 1916 (period II). Then the calculations were repeated for 21 company pairs in two groups of seven firms chosen at random from table A-1. Four pairs of likely competitors were identified with highly correlated residuals: ACF-PSC (I: .425, II: .494); ALOC-PSC (I: .548, II: .565); RSS-PSC (I: .398, II: .389); and IP-UBP (I: .520, II: .552). Note that three of them involve Pressed Steel Car. The other correlation coefficients were usually much smaller, as shown in table A2. The results indicate that residual dependence should not be entirely ignored, but also that it is concentrated in PSC and either IP or UBP.

These two sources of relatively large covariance were eliminated by recomputing the summary statistics of the residual distributions in table 4 for a 26-firm control group, after deleting PSC and either IP or UBP on a random basis. Of course this procedure does not account for all the residual dependence in the sample, but it is preferable to no adjustment whatever, and it tends to raise the standard deviations—as would be expected with a full correction for generally positive covariance. However the increases were not sufficient to alter the confidence levels of any trust-successors residual test in table 4, because the significant entries are themselves quite large in absolute value. Consequently, that table presents the distribution summaries for the entire sample of 28 corporations.

The normality of the control group residual distributions was evaluated with a

Kolmogorov-Smirnov (KS) test for goodness of fit.[34] This test is primitive because the parameters of each distribution were estimated from the data. However, the results strongly supported the assumption of normality: using either the 26- or 28-firm sample, the KS statistic was insignificant at the 95 percent level for 20 of the 21 separate distributions summarized in table 4.

References

Ball, Raymond J. "Changes in Accounting Techniques and Stock Prices." Ph.D. dissertation, Univ. Chicago, 1972.

Blattberg, Robert C., and Gonedes, Nicholas J. "A Comparison of the Stable and Student Distributions as Statistical Models for Stock Prices." *J. Bus.* 47 (April 1974): 244–80.

Burns, Malcolm R. "The Economic Effects of Horizontal Dissolution Decrees under the Sherman Act: A Portfolio Analysis." Ph.D. dissertation, Univ. Minnesota, 1975.

Collins, Norman R., and Preston, Lee E. "The Size Structure of the Largest Industrial Firms, 1909–1958." *A.E.R.* 51 (December 1961): 986–1011.

Commercial and Financial Chronicle. New York: Dana, 1903–17.

Cowles, Alfred. *Common-Stock Indexes.* Bloomington, Ind.: Principia Press, 1938.

Cox, Reavis C. *Competition in the American Tobacco Industry.* New York: Columbia Univ. Press, 1933.

Dewey, Donald. *Monopoly in Economics and Law.* Chicago: Rand McNally, 1959.

Ellert, James C. "Antitrust Enforcement and the Behavior of Stock Prices." Ph.D. dissertation, Univ. Chicago, 1975.

Fama, Eugene F. "The Behavior of Stock Market Prices." *J. Bus.* 38 (January 1965): 34–105.

———. "Efficient Capital Markets: A Review of Theory and Empirical Work." *J. Finance* 25 (May 1970): 383–417.

———. "A Note on the Market Model and the Two-Parameter Model." *J. Finance* 28 (December 1973): 1181–85.

Fama, Eugene F.; Fisher, Lawrence; Jensen, Michael; and Roll, Richard. "The Adjustment of Stock Prices to New Information." *Internat. Econ. Rev.* 10 (February 1969): 1–21.

Fama, Eugene F., and MacBeth, James D. "Risk, Return, and Equilibrium: Empirical Tests." *J.P.E.* 81, no. 3 (May/June 1973): 607–36.

Fama, Eugene F., and Miller, Merton H. *The Theory of Finance.* New York: Holt, Rinehart & Winston, 1972.

Fisher, Lawrence, and Lorie, James H. "Rates of Return on Investments in Common Stocks." *J. Bus.* 37 (January 1964): 1–21.

Hamada, Robert S. "Portfolio Analysis, Market Equilibrium and Corporation Finance." *J. Finance* 24 (March 1969): 13–31.

Hays, William L., and Winkler, Robert L. *Statistics: Probability, Inference, and Decision.* New York: Holt, Rinehart & Winston, 1971.

Jordan, Ronald J. "An Empirical Investigation of the Adjustment of Stock Prices to New Quarterly Earnings Information." *J. Financial and Quantitativ Analysis* 8 (September 1973): 609–20.

Lev, Baruch, and Mandelker, Gershon. "The Microeconomic Consequences o Corporate Mergers." *J. Bus.* 45 (January 1972): 85–104.

Lorie, James H., and Hamilton, Mary T. *The Stock Market: Theories and Evidenc* Homewood, Ill.: Irwin, 1973.

---

[34] The KS test is described in Hays and Winkler (1971, pp. 817–21).

Markowitz, Harry. *Portfolio Selection: Efficient Diversification of Investments.* New York: Wiley, 1959.

Martin, John D., and Klemkosky, Robert C. "Evidence of Heteroscedasticity in the Market Model." *J. Bus.* 48 (January 1975): 81–86.

Modigliani, Franco, and Miller, Merton H. "The Cost of Capital, Corporation Finance and the Theory of Investment." *A.E.R.* 48 (June 1958): 261–97.

Moody's Manual of Railroad and Corporation Securities. New York: Moody, 1903–9.

Nicholls, William H. *Price Policies in the Cigarette Industry.* Nashville, Tenn.: Vanderbilt Univ. Press, 1951.

Ornstein, Stanley I. "Concentration and Profits." *J. Bus.* 45 (October 1972): 519–41.

Pechman, Joseph A. *Federal Tax Policy.* Washington: Brookings Inst., 1966.

Poor's Manual of Industrials. New York: Poor's Manual, 1910–17.

Schall, Lawrence D. "Asset Valuation, Firm Investment, and Firm Diversification." *J. Bus.* 45 (January 1972): 11–28.

Scherer, F. M. *Industrial Market Structure and Economic Performance.* Chicago: Rand McNally, 1970.

Standard Oil Company of N.J. v. U.S., 221 U.S. 1 (1911).

Stigler, George J. "Comments." *A.E.R.* 46 (May 1956): 504–7.

———. "The Dominant Firm and the Inverted Umbrella." *J. Law and Econ.* 8 (October 1965): 167–72.

Stocking, George W. *The Oil Industry and the Competitive System.* Cambridge, Mass.: Riverside, 1925.

Tennant, Richard B. *The American Cigarette Industry.* New Haven, Conn.: Yale Univ. Press, 1950.

Tobin, James. "Liquidity Preference as Behavior towards Risk." *Rev. Econ. Studies* 25 (February 1958): 65–86.

U.S. v. American Tobacco Co., 221 U.S. 106 (1911).

U.S. Bureau of Corporations. *Report of the Commissioner of Corporations on the Petroleum Industry.* Vol. 1. Washington: Government Printing Office, 1907.

———. *Report of the Commissioner of Corporations on the Tobacco Industry.* 3 vols. Washington: Government Printing Office, 1909–15.

U.S. Courts. *Decrees and Judgments in Federal Anti-Trust Cases: July 2, 1890–January 1, 1918.* Washington: Government Printing Office, 1918.

Whitney, Simon N. *Antitrust Policies.* 2 vols. New York: Twentieth Century Fund, 1958.

Williamson, Harold F., and Andreano, Ralph L. "A Reappraisal of the Competitive Structure of the American Petroleum Industry before 1911." In *New Views on American Economic Development*, edited by Ralph L. Andreano. Cambridge, Mass.: Schenkman, 1965.

Capital Punishment and Deterrence: Some Further Thoughts and Additional Evidence

---

Isaac Ehrlich

University of Chicago and National Bureau of Economic Research

Investigation of the deterrent effect of capital punishment has implications far beyond the propriety of execution as punishment since it concerns the general question of offenders' responsiveness to incentives. This study challenges popular allegations by earlier researchers denying the deterrence hypothesis. The empirical analysis based on cross-sectional data from the U.S. corroborates my earlier analysis of the time series. Findings indicate a substantial deterrent effect of punishment on murder and related violent crimes and support the economic and econometric models used in investigations of other crimes. Distinctions between classes of executing and nonexecuting states are also examined in light of theory and evidence.

The importance of determining the deterrent effect on crime of various punishments is one of the few issues concerning deterrence that is not controversial. The problem of deterrence is encountered in all socially disapproved behavior, but it has a natural and dramatic significance with respect to the crime of murder. The present essay will examine the evidence on the cross-sectional patterns of murder and execution in 1940 and 1950, both as an illuminating source of information on the problem

I am indebted to my colleagues Lawrence Fisher, Ronald Gallant, James Heckman, Sam Peltzman, George Stigler, and to two anonymous referees for numerous comments and suggestions. I am particularly grateful to Randall Mark for his general research assistance and suggestions and to Joel Gibbons for useful comments on a previous draft. I also wish to acknowledge the competent computational assistance of Linda Moy and Vasant Sukhatme. I am also indebted to Kenneth R. Brimmer of the National Prisoner Statistics of the Bureau of the Census for his help in obtaining execution data for the years 1936–40. I alone am responsible for errors. Financial assistance for this study has been provided by a grant from the National Science Foundation to the National Bureau of Economic Research for research in law and economics. The financial support of the Charles C. Walgreen Foundation is also gratefully acknowledged. The paper is not an official NBER publication.

[*Journal of Political Economy*, 1977, vol. 85, no. 4]

© 1977 by The University of Chicago. All rights reserved.

of deterrence and as a supplement to my earlier study of the time series.[1] The cross-sectional studies on balance reinforce the findings of the substantial restraining effect on murders of punishment in general and capital punishment in particular. The specific results obtained are consistent with sharp predictions emanating from the deterrence theory.

I deal here with the relationship between estimates of the murder rate and the conditional probability of execution, but the present investigation explores more fully than the previous study the nature of deterrence. The time-series analysis was based on a small number of observations, whereas the present analysis uses a larger body of independent, cross-sectional data relating to different states in the United States. The cross-sectional data include estimates of severity of imprisonment for murder and other crimes, missing from the time series, and provide useful information on additional relevant variables. While the time-series analysis deals with aggregate data relating to both executing and nonexecuting states, the cross-sectional statistics allow separate investigation of the subsets of executing and nonexecuting states and other groupings. The more complete data on punishments imposed for other crimes also permit a partial investigation of the interaction between murder and related violent crimes and thus provide for a somewhat more complete assessment of specific deterrent effects. The cross-sectional data from 1940 and 1950 are advantageous also because they relate to years in which the level of enforcement of capital punishment in executing states appears to have been sufficiently high and variable to permit a meaningful statistical test of the hypothesis of no deterrence. And the general uniformity in the reporting of crime rates in 1940 and 1950 permits some testing of temporal as well as cross-sectional homogeneity. In contrast, however, data exigencies unique to the cross-sectional samples have at this stage precluded utilization of the relatively more sophisticated estimation techniques employed in my earlier investigations.

Some economists may feel that a study of murder is of little relevance for the purpose of testing the economic approach to criminality. Indeed, my main concern is the general question of offenders' responsiveness to incentives, an issue I have investigated in earlier studies by stressing other crimes and different punishments (see Ehrlich 1974). I return to the seemingly narrower study of murder both because of the presumed relative superiority of murder statistics and because I find no compelling reason to expect perpetrators of crimes of passion necessarily to be less responsive to incentives than perpetrators of crimes involving material gains only. Those with a greater proclivity or "taste" for violence are indeed more likely to commit assault or murder than others, given the same objective motives and equal prospects of punishment. However, they need not be less responsive to *changes* in these factors. Inelasticity of demand is not necessarily an increasing function of the extent of demand at given income and prices. Being emotional does not preclude the ability to make self-serving choices. Even if control over actions is foregone temporarily during the commission of crime, the prospect of punishment through public or private sanctions still may prevail upon individuals to avoid situations in which "loss of control" is likely to occur. And if rational people fear death more than other punishment, the death penalty should have the greatest deterrent effect. Conclusions reached in research concerning the abhorrent crime of murder, which some criminologists consider a special form of deviance, thus can shed important light on the relevance of the economic approach to crime and to human behavior in general.

---

[1] Ehrlich (1975*a*, 1975*b*). A more complete critical review of the literature on the deterrent effect of capital punishment will be included in a forthcoming monograph on the topic.

The plan of the paper is as follows: Section I discusses some analytical issues unique to this study. Section II presents an econometric investigation of cross-state variations in murder and related crimes and analyzes the results. Section III examines data on differences in offense rates, conviction risks, and sanction levels across executing and non-executing states and analyzes these data in light of a model of optimal demand for public law enforcement. A brief comparison with previous findings is offered in the concluding section.

I. Analytical Considerations

The general model underlying the research into punishment and deterrence continued in this paper concerns three major structural relations comprising a system of crime and law enforcement: supply of offenses, demand for enforcement activities, and production of means of deterrence and prevention. Data exigencies and other related problems sharply limit at this point the opportunities for an effective investigation of this general system. As did the preceding time-series study of murder, the present investigation focuses on analysis of supply of offenses. The discussion in this section sets the stage for the pursuant empirical investigation. Considerations relevant to demand for enforcement, including imposition of the death penalty, are briefly discussed in Section III.

A. Deterrence Variables as "Prices"

If the decision to commit murder and other related crimes is considered to be the result of an independent either/or choice among alternative actions by a potential offender in a given period, then murder would be expected to occur if the expected utility from committing murder to that person U_m^* exceeds the expected utility from alternative pursuits. Let the outcome of murder consist of the three consumption prospects $C(o)$, $C(m)$, and $C(d)$ relating to three mutually exclusive states of the world: no punishment, entailing no imposed costs; imprisonment, entailing the (discounted) cost of m "dollars"; and execution, entailing the larger cost of d.[2] Assuming that the utility function is bounded, the expected utility from murder to the perpetrator is then given by

$$U_m^* = (1 - Pc)U[C(o)] + (Pc - Pe)U[C(m)] + PeU[C(d)], \quad (1)$$

where $Pc - Pe \equiv Pm \equiv Pc(1 - Pe|c)$ and $Pe \equiv Pc\, Pe|c$ define, respectively, the conditional probabilities of imprisonment and execution (given crime), and $Pe|c$ defines the conditional probability of execution, given conviction. Equation (1) establishes the role of the exhaustive set of deterrence variables Pm, Pe, m, and d, or alternatively, Pc, Pe, m, and d as "prices" or negative rewards for murder. The latter set of prices is of particular interest in this analysis, since the effect of an increase in Pe given Pc measures the differential impact of execution relative to imprisonment rather than that of execution relative to no punishment. Moreover, one may infer not only the signs but also the ranking of the partial elasticities of U_m^* with respect to the *independent* components of Pc and Pe:[3]

$$\varepsilon_{Pc \cdot Pe|c} > \varepsilon_{Pe|c \cdot Pc} \equiv \varepsilon_{Pe \cdot Pc} > 0; \qquad \varepsilon_m > 0, \quad (2)$$

where $\varepsilon_{i \cdot j} \equiv -(\partial U_m^*/\partial i)(i/U_m^*)$, with j and the remaining relevant deterrence variables held constant.[4]

The decision to commit murder may be affected, however, not only by own "prices" but also by the costs and rewards associated with related crimes such as aggravated assault and robbery. These related crimes could be either "substitutes" or "complements" with murder. It is conceivable that an intended act of robbery or assault may result in, and thus be classified as, murder in the event the victim is fatally injured. Let μ be the probability that an intended robbery results in murder and assume, for methodological simplicity, that an offender guilty of murder cannot be convicted of a lesser offense. The expected utility from robbery is then

---

[2] Two additional basic hazards should be recognized in connection with murder and other related crimes One is death through legal intervention by police. The other involves similar detriments through private sanctions such as lynchings. Incorporation of these events formally in equation (1) would not affect the implications stated in equation (2) if the additional events are defined to be independent of events originally summarized in equation (1). Empirically it may be important, however, to control for the probabilities of these other events if their levels are systematically related to those of deterrence and other explanatory variables. While the annual number of reported lynchings in the United States declined to negligible magnitudes after 1935 (see the *Historical Statistics of the United States from Colonial Times to 1957* [U S. Bureau of the Census 1960, p. 218]), the total number of annual homicides inflicted through police intervention has exceeded that of executions in recent decades (282 homicides relative to 81 executions in 1950; 412 homicides relative to no executions in 1971; see *Vital Statistics of the United States* [U.S. Department of Health, Education, and Welfare 1953, 1975]). However, such homicide data are not classified by the type of crime leading to their perpetration, nor are any statewide data available prior to 1949. A test I have conducted in connection with the effect of a measure of the risk of death through police intervention on the rate of violent crimes in 1950 is briefly discussed in Section II (see n. 24).

[3] Focusing on the role of the independent components of deterrence variables has an analytical advantage in that these are produced separately through distinct law enforcement activities. They may thus be thought of as the relevant objects of choice from the viewpoint of the demand for optimal enforcement (see Section III).

[4] The identity $\varepsilon_{Pe|c \cdot Pc} \equiv \varepsilon_{Pe \cdot Pc}$ follows from the definition of Pe as a product of Pc and $Pe|c$. For this reason the theory also implies that $\varepsilon_{Pc \cdot Pe|c} \equiv \varepsilon_{Pc \cdot Pe} + \varepsilon_{Pe \cdot Pc}$.

$$\begin{aligned} U_R^* &= \mu(1 - Pc^{rm})U[C^{rm}(o)] + (1 - \mu)(1 - Pc^r)U[C^r(o)] \\ &\quad + (1 - \mu)Pc^rU[C^r(r)] + \mu Pm^{rm}U[C^{rm}(m)] + \mu Pe^{rm}U[C^{rm}(d)]. \end{aligned} \tag{3}$$

In equation (3) $Pm^{rm} \equiv Pc^{rm}(1 - P^{rm}e|c)$, and $Pe^{rm} \equiv Pc^{rm}Pe|c^{rm}$, denote the probabilities that a robber guilty of robbery-murder will be punished through imprisonment or execution, respectively; Pc^r denotes the probability of conviction of "simple" robbery, for which imprisonment (r) is the only penalty; and C^{mr} and C^r designate the consumption prospects in case robbery results in murder or in a simple robbery only. If μ were fixed by nature and if robbery were not a close alternative to murder as an expression of malevolence, then robbery and murder would be "complements" in the sense that an increase in the set of prices for murder including the risk of execution would have similar deterrent effects on robbery as well.[5] Similarly, an increase in deterrence variables for murder, or a decrease in relative financial gains from robbery, would "deter" robbery. Like considerations apply in connection with assault.

But substitutability relations among murder and robbery or assault can also be predicted on theoretical grounds. Assault may be the closest alternative to murder as an act of malice. To the extent that the two activities are independent ("μ" is subject to control), an increase in the own price of murder might enhance the incentive to "produce" assault. A similar consideration applies in connection with a potential trade-off between the commission of robbery-murder as opposed to a simple robbery. The net effect of a change in, say, the risk of execution for murder on the frequency of robbery and assault must therefore be settled on empirical grounds.

B. The Effects of Other Variables

Whether the expected utility from crime exceeds that from the relevant alternative depends upon the relative rewards as well as the relative prices. In most cases of murder the rewards are expected to be largely psychic, depending on such intangibles as hate or the degree of social interaction (see Ehrlich 1975*a*, pp. 399, 412). By contrast, the cost of imprisonment would be an increasing function of the opportunity cost of time which, for most potential offenders, is derived from legitimate earning opportunities. Moreover, relative legitimate and illegitimate opportunities may affect even the frequencies of murder and assault because of the interdependencies between these crimes and robbery or other crimes against property. Indirect proxies for such economic variables—the level and distribution of income—which were developed and tested in previous investigations (Ehrlich 1974, 1975*a*) will be examined in this analysis as well.

---

[5] The elasticities of U_R^* with respect to Pc^{rm} and $Pe|c^{rm}$ will be "weighted," however, by μ.

The interdependencies among murder, assault, and robbery are determined not only through the effects of prices and relative rewards, however, but also through the influence of medical technology and the cost of saving lives as well as the technology and cost of deadly weapons. These factors are relevant because they can influence the probability that a murderous attack such as aggravated assault results in the death of a victim. The probability of nonsurvival thus determines in large measure the probability that aggravated assault results in murder (the analogue of μ in eq. [3]). That the effect of medical technology on the incidence of murder in different areas over the period under investigation may be significant relative to advances in the technology of deadly weapons is indicated by significant and systematic variations in the ratio of murders to the sum of aggravated assaults and murders $\mu = M/(M + A)$, which is a rough indicator of nonsurvival risks in murderous attacks. The ratio of murders relative to all aggravated assaults and robberies also is presented. Table 1 shows that all rates have decreased between 1940 and 1970 and that nonsurvival risks in urban areas are lower than those in rural areas, although the differences have narrowed over time possibly because of the trends toward equalization of medical and hospital care (see Pashigian 1973). This analysis provides an independent justification for introduction of family income, indicators of relative poverty, and even the degree of urbanization among the set of variables accounting for state-wide variations in murder rates. Clearly, the quality and extent of utilization of medical and hospital care are functions of income and vary systematically across urban and rural areas and different racial groups (see, e.g., Aday and Anderson 1975).

C. The Empirical Supply-of-Offenses Function

The model presented in Part A of this section does not yield definitive implications concerning the form of the individual supply-of-offenses function. The form of the aggregate function, in turn, depends on that of individual functions as well as on the personal distribution of psychic returns for crime if the magnitudes of deterrence variables and other determinants of pecuniary net rewards were perceived to be the same by all persons. Several considerations imply, however, that for the purpose

TABLE 1

ESTIMATES OF NONSURVIVAL RISKS IN MURDEROUS ATTACKS BY POPULATION GROUPS, 1940–70

POPULATION GROUP	1940*		1950†		1960‡		1970§	
	$\frac{M}{M+A}$	$\frac{M}{M+A+R}$	$\frac{M}{M+A}$	$\frac{M}{M+A+R}$	$\frac{M}{M+A}$	$\frac{M}{M+A+R}$	$\frac{M}{M+A}$	$\frac{M}{M+A+R}$
Total cities	.105	.052	.065	.040	.050	.030	.044	.0195
Cities over 50,000	.103	.049	.062	.037	.046	.027	.047	.0193
Cities under 50,000	.113	.064	.076	.053	.070	.048	.030	.0207
Rural areas‖	.163	.108	.138	.098	.134	.105	.066	.058

SOURCES.—U S Department of Justice, *Uniform Crime Reports* (1940, no. 4, tables 76 and 85; 1950, no. 2, table 31; 1951, no 2, p 49; 1960, tables 1 and 6, 1970, tables 1 and 9).

* Total cities' crimes based on 2,001 cities with population of 65,128,946, rural crimes as reported by 1,016 sheriffs, nine state police organizations, and 66 village officers.

† Total cities' crimes based on 2,297 cities with population of 69,643,614; rural crimes as reported by 1,621 sheriffs, 162 rural village officers, and 11 state organizations representing a rural population of 42,433,145

‡ Total cities' crimes based on 3,366 cities with population of 96,678,066, rural crimes are estimated totals in the U S. rural population of 41,832,427

§ Total cities' crimes based on 4,481 cities with population of 121,254,000, rural crimes are estimated totals in the U S. rural population of 38,865,034.

‖ Statistics on rural crimes in 1970 and 1960 may not be fully comparable to those in 1950 and 1940 The latter may have been limited in some instances only to cases in which arrests were made.

of a simple empirical formulation a linear specification in the logarithms of the dependent and basic independent variables is superior to a linear specification in the natural values of these variables. In the first place, the theory identifies the basic deterrence variables in the murder equation to include the *unconditional* probabilities of conviction and execution which are functionally dependent via a multiplicative relationship $Pe \equiv Pc\, Pe|c$. Introduction of natural values of these variables in a regression analysis is then likely to produce imprecise estimates of their independent effects. The effect of the independent variable $Pe|c$, in turn, depends on the level of Pe in the same way that the effect of an ad valorem tax rate depends on the level of the unit price. A linear specification in Pc and $Pe|c$ is disadvantageous since it does not control directly for the level of the basic price Pe. In contrast, a linear specification in the logarithms[6] permits an equally effective estimation of the deterrent effects of interest via the following equation:

$$\ln q = \alpha - \beta_1 \ln Pc - \beta_2 \ln Pe|c \ldots, \qquad (4)$$

---

[6] The partial elasticities of the expected utility from crime of an average potential offender, $\bar{U}^*$, with respect to Pc and $Pe|c$, for example, can be shown to be an increasing function of these variables. However, the supply of offenses relates to the average flow of crimes in the population, q, rather than to $\bar{U}^*$; the latter two variables are monotonically, but not necessarily proportionally, related. Formally, given that $dq/d\bar{U}^* > 0$, the effects of deterrence variables on q may be summarized by $e_p = e_{q\bar{U}^*} \varepsilon_p$, where $e_p = -(\partial q/\partial p)(p/q)$, $e_{q\bar{U}^*} = (dq/d\bar{U}^*)(\bar{U}^*/q)$, $\varepsilon_p = -(\partial \bar{U}^*/\partial p)(p/\bar{U}^*)$, and p stands for Pc or $Pe|c$. Clearly, the value of $\partial e_p/\partial p$ cannot be inferred directly from that of $\partial \varepsilon_p/\partial p$, since the theory does not provide clear insights as to the way $e_{q\bar{U}^*}$ changes with $\bar{U}^*$. The assumption that the relevant e's are constant can thus be taken as a useful first approximation. Also see Becker (1968, p. 209).

or the statistically identical form

$$\ln q = \alpha - \gamma \ln Pc - \beta_2 \ln Pe, \tag{4a}$$

where $\gamma = \beta_1 - \beta_2$.[7] The logarithmic specification thus controls for the magnitudes of the basic prices *Pc* and *Pe* and at the same time eliminates their functional interdependence and permits an efficient test of the hypotheses summarized in equation (2).

Furthermore, the logarithmic form is a superior specification when the magnitudes of errors in the dependent variable are thought to be proportional to its level. It would be rather implausible to assume that the magnitudes of underreported and misreported crimes are independent of the level of reported crimes. It would be more plausible to expect a proportional relationship between the two. Similar considerations also apply to the observed values of deterrence variables (see Appendix B). This assumed proportionality of errors to levels of corresponding variables implies that the logarithmic specification is at least tractable in terms of an analysis of expected biases (see Ehrlich 1974, appendix I in his section IV).

These arguments can be evaluated systematically by application of a test of functional specification devised by Box and Cox (1964). The test provides a systematic method of determining an optimal transformation of the dependent variable or the entire set of variables entering a regression equation (X_i) of the form $(X_i^\lambda - 1)/\lambda$. This class of transformations is directly applicable in the present context, in that $\lambda = 0$ defines the logarithmic transformation and $\lambda = 1$ corresponds to the simple-linear specification. A related class of transformations permits a comparison between the logarithmic and the semilogarithmic functional forms. The assumption underlying Box and Cox's analysis is that some transformation λ exists such that the transformed observations satisfy the full normal theory assumptions and are linearly related to their determinants. The procedure then permits selection of an optimal transformation and testing of hypotheses concerning alternative values of λ as optimal transformations. A comprehensive application of this procedure is reported in Section IIB2 and Appendix A.

II. Empirical Implementation

A. Introduction

The supply-of-offenses functions relating to murder and other crimes are estimated in this investigation using cross-state data from 1940 and 1950. These years are chosen because the level of enforcement of capital punishment in individual executing states appears to have been sufficiently high and variable to permit a meaningful test of the hypothesis of no deterrence. Due to data exigencies, the parameters of the supply functions are estimated via classical least-squares and seemingly unrelated simultaneous-equations techniques rather than the two-stage least-squares procedure employed in my related works. Because the preferred regression format is found to be logarithmic and the estimated levels of conditional risks of execution in nonexecuting states are "zero," supply-of-offenses functions are estimated separately in executing states as well as in the full samples comprised of executing and nonexecuting states. In the full-sample regressions estimated via the approximate logarithmic transformation, statistical considerations dictate allowance for separate constant terms in each class of states. Following tests for homoscedasticity, generalized least-squares estimates are achieved by weighting all variables by the square roots of either urban, state, or (when available) the relevant sample populations. A detailed discussion of the estimation procedure and the variables used in the regression analysis is included in Appendix B. A list of the variables used is given in table 2.

---

[7] This result, which is "forced" by the logarithmic specification, is, of course, consistent with the theoretical analysis itself (see n. 4 above).

B. Analysis of Regression Results

1. Analysis of FBI (UCR) Data

FBI statistics on murder and other felonies are available in 1940 and 1950 only as crime rates in samples of urban areas covering, on average, 49 percent of the states' residential population in 1940 and 45 percent in 1950 (see U.S. Department of Justice, *Uniform Crime Reports*, 1940, 1950–51, 1960, 1970). Crime statistics are based on complaints of crime, subject to police verification procedures, and are recorded according to *Uniform Crime Reports* (*UCR*) classifications. The category "murder" refers to all willful felonious homicides.

a) *Subsets of states with positive executions.*—The basic regression model used in the analysis is linear in the logarithms of all relevant variables:[8]

$$E(\ln q^o) = a + b_1 \ln T + b_2 \ln P^o c + b_3 \ln P^o e|c + b_4 \ln NW + b_5 \ln X + b_6 \ln W + b_7 \ln \mathrm{AGE} + b_8 \ln \mathrm{URB}, \qquad (5)$$

---

[8] The set of variables included in equation (5) generally yielded consistent and significant results in the different regressions performed. Other variables tested are measures of unemployment and labor-force participation rates. As in my previous cross-sectional investigations, these variables have not been found to have a systematic impact on the extent of crime rates across states, although both variables did appear to have such effects in my related time-series analysis (see the discussion in Ehrlich [1975*a*]). Their exclusion from the regression equations did not have any significant impact on the regression results reported here.

TABLE 2

LIST OF VARIABLES USED IN THE REGRESSION ANALYSIS

q_i^o = urban crime rate (per 100,000 population) for offense category i, based on *UCR* samples;

h^o = homicide rate (per 100,000 population), based on *VS* data;

$P^o c_i = \frac{C_i^o}{Q_i^o}$ = estimate of probability of conviction: the ratio of prisoners received in state prisons to the estimated (state) total numbers of offenses in category i;

$P^h c = \frac{C^o}{H^o}$ = corresponding estimate of probability of conviction for murder using *VS* data on the number of homicides in a state;

T_i† = median time spent in state prisons by offenders prior to first release;

$PX5 = \frac{\sum_{j=0}^{4} E_{t-j}^o/5}{C_t^o}$ = the ratio of the average number of executions in years t to $t-j$ to the estimated number of convictions for murder in sample year t;

$PX4 = \frac{\sum_{j=0}^{3} E_{t-j}^o/4}{C_t}$ $\quad PXQ1 = \frac{E_{t+1}^o}{C_t^o}$ $\quad PXQ2 = \frac{E_t^o}{C_t^o}$ $\quad PXA = \frac{E_{t+1}^o + E_t^o/2}{C_t}$

$PXQ1Q - PX5Q$‡ = same as the preceding ratios but with E_t^o defined to include only executions for murders;

W§ = median income of families in year $t - 1$: a census estimate;

X = percentage of families with income below one-half of W;

NW = percent of nonwhites in the population;

AGE = percent of population in the age group 15–24;

A_{25-34} = percent of population in the age group 25–34;

URB = percent of urban population in the total state population;

EPOS = a dummy variable distinguishing executing states (1) from states (0) with $P^o e | c = 0$;

N_s = state population in 100,000s;

N_{URB} = urban population in 100,000s;

N_{REP} = size of population (in 100,000s) in *UCR* samples reporting q_i^o in 1951;

NOTE *Due to data exigencies, UCR murder regressions omit from the full samples in 1940 and 1950 the* following states 1940 Georgia, Maine, Mississippi, Nevada, and Vermont, 1950– Georgia, Michigan, North Dakota, and Vermont

† In the 1940 regressions data exigencies dictated the use of the corresponding variable from 1950 (T_{50}). The actual data are for prisoners released in 1951 (1952 for T_{MUR} in Idaho, Michigan, Rhode Island, South Dakota, and Vermont) because 1950 figures are not available

‡ In 1940 $PX5Q$ cannot be computed because execution data for 1936 are not available by crime category

§ In 1940 available only as median earnings of a sample of wage and salary workers

with all variables weighted by the weighting factors discussed in Appendix B.[9] Table 3 presents estimates of regression coefficients and the ratios of these coefficients to their standard errors. Magnitudes of the "R^2"

---

[9] According to the discussion in Section IIA and Appendix B, weighted regressions are performed to derive generalized-least-squares (GLS) estimates of equation (5). *Indeed, the standard errors of regression estimates derived from weighted regressions are* virtually with no exception lower than the standard errors of corresponding unweighted regressions, a sample of which is given in table 3. The fact that weighted regressions using the square roots of urban and reporting populations as weighting factors are found to be somewhat more efficient than those derived from regressions weighted by the square root of the state population also is consistent with the heteroscedasticity argument, since the dependent variables are estimates of urban crime rates.

statistics reported in connection with weighted regressions are the conventional R^2 measures obtained after multiplying the variables in equation (5) by a weighting factor. These "R^2" measures thus may serve merely as indicators of goodness of fit in the transformed regression equations.

Results obtained in connection with the three deterrence variables, T, P^oc, and measures of $P^oe|c$, exhibit a consistent pattern throughout the regression analysis. The estimated regression coefficients associated with these variables are all negative, with magnitudes generally well above twice the values of corresponding standard errors. More important, with no exception, the estimated elasticities of the murder rate, with respect to the probability of conviction (P^oc), exceed those with respect to the conditional probability of execution, given conviction ($P^oe|c$). In all of the cases reported in tables 3 and 4, the differences between the magnitudes of the latter elasticities are significant at the 5 percent level for a two-tail test.[10] As for the effect of T, it should be recalled that the variable used in both the 1940 and the 1950 regressions is the median length of time served by those released from prisons in 1951. In spite of potential biases which may be injected into regression estimates in 1940 as a result, estimates of deterrence effects are statistically indistinguishable across different regressions and do not depend on the inclusion of T_{50} in the 1940 regressions (see Part 5 below).

In contrast to time-series estimates of the simple correlation between measures of the murder rate and the conditional probability of execution, these simple correlations are found to be negative in all the relevant subsets of executing states.[11] In addition, estimates of deterrent effects derived from regressions in which only the three deterrence variables, T, P^oc, and $P^oe|c$, served as regressors show similar qualitative results, with estimates of deterrent effects actually larger in absolute magnitudes.

Except for estimates derived from unweighted regressions in 1950, virtually all 95 percent confidence intervals associated with the estimated

---

[10] All tests of significance reported in this investigation are conditional on the assumption that the error terms satisfy all the standard conditions of the regression model.

[11] The simple (unweighted) correlation coefficients between natural values of q^o and the measures of the conditional risk of execution listed below are:

1940	1950
$PXQ_1 = -.13697$	$PXQ_1 = -.11230$
$PXQ_2 = -.27668$	$PXQ_2 = -.17299$
$PX5 = -.18251$	$PX5 = -.00605$
$PX4Q = -.24158$	$PX5Q = -.10576$

The simple regression coefficients associated with these same measures of execution risk in weighted logarithmic regressions were always negative, and higher in absolute value than the coefficients reported in table 3 and n. 12. Note that the cross-state measures of the risk of execution given conviction of murder are based on independently reported conviction (prison commitments for both imprisonment and execution) data, which do not depend on the FBI estimates of murder rates.

TABLE 3

Murder Supply Functions: GLS (Weighted) and OLS Regressions, Logarithmic Form

Eq.	Year	N	"R^2"	Constant	T	P^o_c	$P^o_{e\|c}$	NW	X	W	AGE	URB
Weight = $\sqrt{N_{URB}}$:												
							$PX5$					
1.	1940	33	.9539	7.372 (1.89)	−0.292 (−2.35)	−0.673 (−4.70)	−0.339 (−3.48)	0.508 (7.26)	. . . . . .	0.436 (1.21)	−2.101 (−2.23)	−0.982 (−4.43)
							$PX4Q$					
2.	1940	33	.9565	5.466 (1.44)	−0.232 (−1.93)	−0.709 (−4.99)	−0.389 (−3.79)	0.468 (6.58)	. . . . . .	0.505 (1.44)	−1.741 (−1.92)	−0.993 (−4.63)
Weight = $\sqrt{N_s}$:												
							$PX5$					
3.	1940	33	.9518	5.757 (1.37)	−0.320 (−2.45)	−0.682 (−4.44)	−0.366 (−3.46)	0.487 (6.20)	. . . . . .	0.480 (1.26)	−1.691 (−1.58)	−0.928 (−4.01)
							$PX4Q$					
4.	1940	33	.9585	4.259 (1.09)	−0.249 (−2.03)	−0.725 (−5.07)	−0.430 (−4.23)	0.438 (5.79)	. . . . . .	0.469 (1.33)	−1.393 (−1.41)	−0.882 (−4.08)
Weight = $\sqrt{N_{REP}}$:												
							$PX5$					
5.	1950	35	.9581	−4.622 (−1.17)	−0.459 (−3.59)	−0.687 (−5.23)	−0.272 (−4.74)	0.496 (7.13)	1.200 (2.22)	1.226 (2.65)	0.274 (0.44)	−0.742 (−2.08)

							PX5Q					
6............	1950	35	.9588	−4.690 (−1.20)	−0.422 (−3.35)	−0.670 (−5.15)	−0.270 (−4.81)	0.487 (7.01)	1.143 (2.12)	1.256 (2.74)	0.177 (0.29)	−0.783 (−2.21)
Weight = $\sqrt{N_i}$:												
							PX5					
7............	1950	35	.9503	−3.024 (−0.84)	−0.507 (−3.87)	−0.719 (−5.30)	−0.324 (−5.04)	0.468 (6.23)	0.904 (1.52)	0.921 (1.99)	0.374 (0.63)	−0.703 (−1.94)
							PX5Q					
8............	1950	35	.9512	−3.101 (−0.87)	−0.460 (−3.53)	−0.694 (−5.18)	−0.324 (−5.13)	0.456 (6.13)	0.834 (1.41)	0.940 (2.06)	0.282 (0.48)	−0.738 (−2.05)
Unweighted:												
							PX4Q					
9............	1940	33	.9087	5.633 (1.19)	−0.152 (−0.98)	−0.792 (−6.11)	−0.509 (−5.48)	0.312 (4.89)	... ...	−0.064 (−0.18)	−1.062 (−0.80)	−0.705 (−2.79)
							PX5Q					
10...........	1950	35	.8394	−2.201 (−0.50)	−0.490 (−2.39)	−0.595 (−3.33)	−0 145 (−1.38)	0.460 (4.40)	0.945 (1.07)	0.793 (1.32)	0.192 (0.37)	−0.364 (−0.74)

NOTE.—$\hat{\beta}/S_{\hat{\beta}}$ in parentheses.

elasticities of the murder rate with respect to alternative measures of the conditional risk of execution include only negative values. Nevertheless, the quantitative differences observed among the point estimates of elasticities associated with other measures of $P^{o}e|c$ (not reported here due to lack of space) indicate the inherent limitations of these measures, resulting from the rather sizable relative fluctuations in the small numbers of executions carried out in a given state over successive years.[12] Indeed, the reported measures of $P^{o}e|c$ based on 4- or 5-year averages of the yearly number of executions generally yield more consistent and pronounced results across the 1940 and 1950 samples. The apparent superiority of $PX4Q$, $PX5Q$, and $PX5$ as measures of execution risk may also be due to the fact that the regressions utilizing these variables include a larger number of observations. Thus the results are based on a larger number of degrees of freedom.

The estimated effects of measures of income (W) and income inequality (X) may reflect the influence of different and possibly even conflicting factors.[13] The variables W and X may serve as indicators of the potential gains and the differential gains from property crimes relative to legitimate earning activities (see Ehrlich [1974, section IIIB] and the results in connection with robbery and assault in tables 6 and 7). The generally positive effects of these variables on the murder rate may be due to the fact that murder is often a by-product of these crimes. The direct partial effects of W and X on the incentive to commit acts of malice are more ambiguous, however (see Ehrlich 1975*a*). Presumably, the incentive effects of wealth on crime outweigh the effect of medical care available to victims of aggravated assault. Note, however, that the observed positive effects of X and NW on q^{0} may partly be due to the effect of inferior medical care services available to relatively low-income persons.

The percent of urban population in the total state population (URB)

---

[12] A sample of results associated with alternative measures of $P^{o}e|c$ defined in table 2 substituted for the measures used in equations 1–2 and 5–6 in table 3 relating to 1940 and 1950, respectively, is as follows ($\hat{\beta}/S_{\hat{\beta}}$ in parentheses):

Measure	PXQ_1	PXQ_2	PXA	PXQ_1Q	PXQ_2Q	$PXAQ$
1940 (set 1–2) .	− .132	− .212	− .160	− .118	− .267	(− .151)
	(−2.42)	(−3.36)	(−2.62)	(−2.02)	(−4.03)	(−2.39)
1950 (set 5–6) . . .	− .152	− .335	− .332	− .177	− .327	− .340
	(−2.05)	(−2.92)	(−4.50)	(−2.25)	(−2.25)	(−4.70)

The magnitudes of estimates of other deterrence variables do not change appreciably and remain statistically significant as alternative estimates of $P^{o}e|c$ are substituted. However, standard errors tend to be lowest with the measures presented in table 3.

[13] In the 1940 regressions, X_{40} yielded insignificant and inconsistent results across equations including a different measure of $P^{o}e|c$. However, both W_{40} and X_{40} are based on a sample of employed wage and salary workers with positive earnings in 1939 rather than on a sample of family income in 1949, including families with no income, from which W_{50} and X_{50} are computed. The exclusion of X_{40} from the regression equations had little impact on the estimated magnitudes of other regressors.

is found to be negatively associated with the (urban) murder rate in both 1940 and 1950. The result might suggest that the magnitudes of some explanatory variables computed from statistics relating to the entire state population differ by constant proportions across urban and rural areas, and, consequently, that URB serves as a "correcting factor." The net effect of URB then would not be clear a priori since it would depend on the direction of specific differences. Alternatively, the degree of urbanization of a state may serve as a proxy for other relevant factors, such as access to relevant medical services, the partial effect of which on the murder rate is expected to be negative.[14]

The percent of young age groups (15–24) in the state population (AGE) is not always found to have a systematic association with the murder rate, whereas the effect of the percent of nonwhites in the population (*NW*) is generally positive and significant in all regressions. As in the case of URB, AGE may perform the role of a "correcting factor" which allows derivation of unbiased estimates of the effects of deterrence variables. Presumably, magnitudes of these variables differ across different age brackets because of the differential treatment of young offenders under the law. While this consideration alone might lead to the expectation that q^o and AGE would be positively correlated, other differences in determinants of murder and acts of malice across different age groups may give rise to different associations. According to information compiled from the *Vital Statistics of the United States* (see Klebba 1975), the homicide victimization rate for nonwhites and for the total population has tended to peak at the age group 25–34 (A_{25-34}) in the period under study. Indeed, estimates of equation (5) with A_{25-34} replacing AGE reveal a positive association between A_{25-34} and the murder rate, especially in 1950, contrary to the inconclusive effect of AGE in these regressions. Generally, estimates of deterrent effects in (5) obtained with A_{25-34} as a measure of age composition did not differ markedly from the estimates discussed in the preceding analysis: the effects of $P^{o}c$ and $P^{o}e|c$ were found somewhat higher in absolute values, and that of T somewhat lower. The main change was observed in 1950 regressions in connection with the effect of the percent of nonwhites in the population, which decreased in magnitude although it remained statistically significant.[15] It thus appears that the positive and significant association of NW and q^o may be, at least in part, the result of its relatively high partial correlation with A_{25-34}. However, homicide victimization data reveal that nonwhites face a much higher risk of victimization than do their white counterparts in all age and sex groups. This evidence is not inconsistent with the hypothesis that the higher frequency of murder among nonwhites is the result of relatively low general (and age-specific) earning opportunities which reduce potential incarceration costs and access to relevant medical services at all ages.[16]

---

[14] This tentative interpretation is not incompatible with the fact that URB is found to be negatively associated also with a total homicide measure based on state statistics, especially in 1940 (see table 5). Differences in availability of medical services across urbanized and less-urbanized states may have been particularly high in 1940.

[15] For example, when A_{25-34} replaces AGE in equation 7 of table 3, the nonwhite coefficient drops by nearly half and the effect of W becomes insignificant. However, substituting A_{25-34} for AGE in the 1940 murder regressions does not have any marked effects on other variables' coefficients. Using A_{25-34} in the murder regressions yields somewhat higher "R^2" values than AGE in 1950 and slightly lower ones in 1940. Also, substituting A_{25-34} in the 1950 robbery and assault regressions (e.g., eqq. 2 and 3 of table 6 and in table 7) hardly affects the robbery regression, while the assault equation is affected much the same way as were the (1950) murder regressions.

b) *Full-sample results.*—The basic regression equation estimated using the full set of observations available in 1940 and 1950 is a modified version of the regression model specified in equation (5),

$$E[q^{o(\lambda)}] = a + b_1 T^{(\lambda)} + \cdots + b_8 \mathrm{URB}^{(\lambda)} + b_9 \mathrm{EPOS}, \qquad (6)$$

where EPOS is a dummy variable distinguishing the relevant executing (1) and nonexecuting (0) states, and the transformation (λ) is defined in Section IC. Because the values of measured execution risks in nonexecuting states are zero, the log-linear regression format specified in equation (5) is here approximated by setting $\lambda = 0.001$. The results for the simple linear and the approximate log-linear transformations are reported in equations 1b, 1c, 2b, and 2c in table 4, parts A and B.

The full-sample results which allow for separate constant terms in executing and nonexecuting states via EPOS, especially those derived through the approximate logarithmic transformation, are essentially the same as the corresponding results based on subsamples of executing states. The point estimates of the elasticity of the murder rate with respect to estimates of $P^o e|c$ are virtually identical. More important, application of Chow's test of equality between the comparable regression coefficients associated with variables entering the approximate logarithmic analog of equation (5) in executing and nonexecuting states, except for the constant terms of the separate equations, shows that the hypothesis of equality for that set of regression coefficients cannot be rejected at the 5 percent, and generally much higher, level of significance. This is evident in tests relating to both the 1940, 1950, and pooled samples (see below) and for regressions using alternative measures of execution risk or different weighting factors.[17]

The effect of the dummy variable EPOS, in turn, is found to be significant in all the regressions using the approximate logarithmic transformation and in the weighted regressions using the simple linear specifications. Indeed, the elimination of EPOS from the regression equation is found to have a drastic effect on the standard error and even the sign of the coefficient associated with $P^{o}e|c$ measures. In the approximate log specification of equation (6) the signs of the latter are positive and sometimes significant with magnitudes very nearly zero, whereas in the simple linear specification the coefficients have the expected negative sign and are occasionally significant as well.[18] The technical explanation is straightforward: EPOS is very highly correlated with $P^{o}e|c$, extremely so in the approximate logarithmic versions of equation (6). Thus, its absence from the regression equation produces the expected "specification bias" resulting from a missing variable, and the magnitude of the quantitative change in the coefficient associated with $P^{o}e|c^{(\lambda)}$ is expected to equal the product of the true partial regression coefficient associated with EPOS and the relevant partial regression coefficient relating EPOS to $P^{o}e|c^{(\lambda)}$. Indeed, it is estimated that the extent of "specification bias" resulting from the omission of EPOS is likely to outweigh the magnitude of the true partial coefficient associated with measures of $P^{o}e|c^{(\lambda)}$ in the approximate logarithmic regressions but not in most of the simple linear runs.[19] It is important to stress, however, that the *magnitude* of the coefficient associated with EPOS is an arbitrary number by definition. It is but an artifact of two related factors: the transformation applied to the dependent and independent variables in the regression and the range of the dummy variable measuring EPOS. If the value of EPOS is transformed, for example, by the same value of λ (0.001) used to transform all other variables in the approximate logarithmic regressions, then the quantitative magnitude of the coefficient associated with EPOS in these regressions would be reduced by precisely 1,000-fold. The only valid inference to be drawn from the estimated effect of EPOS is that it indicates the existence of a statistically significant difference between the mean rates of murder in executing and nonexecuting states after the effects of other variables entering equation (6) have been accounted for. This does

---

[16] For some initial attempts to arrive at a methodology for testing for racial differences in criminal activity see Ehrlich (1970). A fuller investigation requires data concerning the level of law enforcement activity directed against specific groups in connection with crimes against the person and efficient measures of group-specific earning (and medical care) opportunities.

[17] For example, the F-statistics based on the sum of squared residuals estimated from equations 1 and 2 in table 8 (with $\lambda = 0.001$), its counterpart in nonexecuting states, and equation 4 in table 8 are $F_{5,74} = 0.686$ and $F_{6,72} = 1.260$, respectively.

---

[18] Estimates of the elasticity of the murder rate with respect to measures of execution risk in equations 1b, 1c, 2b, and 2c in table 4, parts A and B, which were estimated without introducing EPOS in the regression equation ($\hat{b}_e$), are reported in their respective order: 1940: 0.0003 (2.28), −4.491 (−0.88), 0.0006 (3.23), −6.963 (−1.84); 1950: 0.0003 (1.11), −17.35 (−2.58), 0.0001 (0.59), 3.124 (0.49). The numbers in parentheses are the ratios of the coefficients to their respective standard errors.

[19] The estimated regression coefficient associated with $P^{o}e|c$ in the regressions excluding EPOS $\hat{b}_e$ would be given by plim $(\hat{b}_e) = b_3 + b_9 b_{93..}$ where b_3 and b_9 are the partial (population) regression coefficients associated with $P^{o}e|c$ and EPOS in equation (6) and $b_{93...}$ is the partial (population) regression coefficient associated with $P^{o}e|c$ in a regression of EPOS on $P^{o}e|c$ and all other regressors entering (6). Since b_3 is expected to be negative, and b_3 and $b_{93..}$ are found to be positive, the sign of plim $(\hat{b}_e)$ would reverse in those transformations where the term $b_9 b_{93.}$ outweighs the absolute magnitude of b_3. Indeed, LS estimates of $b_{93..}$ verify that $\hat{b}_9 \hat{b}_{93.}$ tends to exceed $|\hat{b}_3|$ in the approximate log-linear transformations but that it falls short of $|\hat{b}_3|$ in the simple linear transformations, with $\hat{b}_3 + \hat{b}_9 \hat{b}_{93..}$ virtually identical to $\hat{b}_e$.

TABLE 4

A. MURDER SUPPLY FUNCTIONS: 1940, 43 STATES

Eq.	λ	$L(\lambda)$*	Constant	T	$P^{o}c$	$P^{o}e\|c$	NW	W	AGE	URB	EPOS
						$\overline{PX5}$					
1a......	$0.100 = \hat{\lambda}$	−5.0524	0.918	−0.231	−0.885	−0.481	0.531	0.370	−1.510	−0.756	4.217
			(0.268)	(−3.026)	(−5.645)	(−3.920)	(9.332)	(2.032)	(−2.057)	(−5.212)	(4.257)
1b......	0.001	−6.2315	−331.71	−0.324	−0.664	−0.338	0.535	0.635	−1.888	−0.963	337.16
			(−3.706)	(−3.014)	(−5.482)	(−3.764)	(9.367)	(2.030)	(−2.160)	(−5.032)	(3.768)
1c......	1.0	−47.1601	−10.22	−0.017	−13.184	−9.992	0.314	−0.002	−0.011	−0.083	3.313
			(−0.806)	(−1.742)	(−2.898)	(−1.888)	(3.880)	(−0.365)	(−0.022)	(−2.252)	(2.421)
						$\overline{PX4Q}$					
2a......	$0.175 = \hat{\lambda}$	−13.6604	−3.538	−0.149	−1.013	−0.777	0.391	0.138	−0.046	−0.414	3.680
			(−0.700)	(−1.805)	(−5.285)	(−4.470)	(5.643)	(0.849)	(−0.044)	(−2.602)	(5.333)
2b......	0.001	−15.4821	−426.98	−0.254	−0.624	−0.429	0.358	0.212	−0.192	−0.529	428.44
			(−4.081)	(−1.665)	(−4.994)	(−4.106)	(5.304)	(0.535)	(−0.130)	(−2.121)	(4.112)
2c......	1.0	−46.9512	−12.76	−0.019	−9.272	−9.363	0.265	−0.002	0.310	−0.068	4.246
			(−0.933)	(−1.892)	(−2.949)	(−2.628)	(2.949)	(−0.426)	(0.453)	(−1.502)	(2.820)

B. Murder Supply Functions: 1950, 44 States

Eq.	λ	$L(\lambda)$*	Constant	T	P^*c	$P^*e\|c$	NW	X	W	AGE	URB	EPOS
						$\overline{PX5}$						
1a........	0.225 $= \hat{\lambda}$	−2.3101	−5.088	−0.149	−0.975	−0.776	0.410	2.905	0.193	0.726	−0.151	2.285
			(−1.722)	(−2.790)	(−4.792)	(−4.849)	(6.405)	(2.689)	(1.708)	(1.623)	(−0.878)	(5.009)
1b........	0.001	−4.5945	−278.12	−0.294	−0.538	−0.274	0.445	1.545	0.881	0.853	−0.342	273.11
			(−4.421)	(−2.439)	(−4.588)	(−4.309)	(6.118)	(2.711)	(1.811)	(1.374)	(−1.020)	(4.313)
1c........	1.0	−25.3507	−4.491	−0.013	−7.475	−25.780	0.207	41.665	0.001	0.485	0.014	3.292
			(−0.291)	(−2.419)	(−3.017)	(−3.903)	(4.287)	(2.234)	(0.466)	(1.703)	(0.444)	(3.119)
						$\overline{PX5Q}$						
2a........	0.15 $= \hat{\lambda}$	−12.5914	−1.481	−0.152	−0.568	−0.268	0.381	1.969	0.151	0.375	−0.022	1.334
			(−0.386)	(−1.412)	(−2.885)	(−1.289)	(3.893)	(1.387)	(0.645)	(0.850)	(−0.084)	(1.421)
2b........	0.001	−13.2404	−157.87	−0.239	−0.401	−0.158	0.375	1.280	0.264	0.403	−0.004	157.57
			(−1.358)	(−1.299)	(−2.941)	(−1.364)	(3.759)	(1.432)	(0.451)	(0.753)	(−0.011)	(1.365)
2c........	1.0	−32.9919	27.257	−0.015	−4.298	0.478	0.147	40.613	0.000	0.301	0.030	1.238
			(1.192)	(−2.109)	(−1.763)	(0.069)	(2.336)	(1.264)	(0.071)	(1.188)	(0.721)	(0.996)

Note.—GLS and OLS regressions, selected transformations of linear form, results from Box and Cox tests on full observation set ($\hat{\beta}/S_{\hat{\beta}}$ in parentheses). Eqq 1a–1c in part A are weighted by $\sqrt{N_{\mathrm{URB}}}$, eqq. 1a–1c in part B are weighted by $\sqrt{N_{\mathrm{REP}}}$; eqq 2a–2c are unweighted throughout

* $L(\lambda)$ is the magnitude of the estimated log-likelihood function at the given value of λ, as defined in eqq. (A2) and (A4) in Appendix A.

not mean, necessarily, that the mean rate of murder in nonexecuting states is lower than the rate in executing states where the level of $P^{o}e|c$ is lowest. More important, as the analysis in Appendix B indicates, it cannot be determined whether the observed significant difference between the estimated constant terms in equations relating to executing and nonexecuting states is the result of "missing variables" that differ significantly across these subsets of states or because of the arbitrary values assigned to measures of the perceived risk of execution in nonexecuting states: if the true risks were higher than "zero" in these states, then the assignment of zero values to estimates of $Pe|c$ will necessarily bias the constant terms of the corresponding regression equations toward negative values, especially in the approximate logarithmic format. That the effect of EPOS sometimes is found to be statistically insignificant in regressions based on the simple linear specification indicates the relative importance of this consideration (see also n. 27).

2. Box and Cox Tests for Optimal Transformations

Although the qualitative results derived through the GLS estimation procedure are found to be robust with respect to functional specification (see table 4, parts A and B), the Box and Cox tests for optimal transformations reported in Appendix A unambiguously establish the relative superiority of the logarithmic specification. In virtually all tests conducted, the logarithmic specifications cannot be rejected as optimal within the classes of power transformations considered. In contrast, the simple linear transformations and the semilogarithmic transformation considered are rejected as optimal in all cases. These results are consistent with the analytical considerations raised in Section IC and lend support to the results on deterrence derived through the logarithmic specification in earlier studies (Ehrlich 1974, 1975*a*). Moreover, the GLS results obtained through the *optimal* transformations ($\hat{\lambda}$) selected by the Box and Cox procedure are without exception compatible with the basic deterrence hypotheses concerning the effects of the probability of conviction, the conditional probability of execution, and the severity of imprisonment.

3. Vital Statistics Data on Homicide

The *Vital Statistics of the United States* (*VS*) (U.S. Bureau of the Census 1943; U.S. Department of Health, Education, and Welfare 1953) provides alternative data on homicides. For the purpose of an empirical implementation of the theory, however, the FBI data are conceptually superior to the *Vital Statistics* series because the latter includes justifiable and excusable homicides, whereas the FBI murder category is defined to include only willful felonious homicides.[20] The *VS* data may also understate the true number of homicides, because death certificates on which these statistics are based sometimes reflect classification of death by medical causes alone, especially in those instances when death from homicidal assaults lags the actual time of assault. The cross-sectional information in the *Vital Statistics* in 1950 is also somewhat noncomparable because in 1940 homicide statistics are reported by the place of the occurrence of homicide, which is the desired statistic, whereas in 1950 these statistics are reported by the place of the victim's residence. Also, only the 1950 data permit homicides caused by the legal intervention of police to be separated from the total homicide counts. One clear advantage of the *VS* data over the *UCR* statistics, however, is that the *VS* information is based on a count of homicides in the entire state, whereas the FBI data are based on samples of urban areas.

The regression models applied to the *VS* data are the models summarized in equations (5) and (6) with the following modifications: the dependent variable now is the homicide rate defined as $h^o = (H^o/N_S)$, where H^o stands for the *VS* total homicide counts[21] and N_S is the state population. The conviction risk is now defined as $P^h c = (C^o/H^o)$. A sample of the regression results is summarized in table 5.

The results are qualitatively the same as those discussed in connection with the FBI data. The quantitative results differ mainly in connection with the estimated deterrent effect of the probability of conviction measures used in the separate analyses which is here smaller in magnitude and less precise. This result is not entirely surprising given that the *VS* data on homicide may not be as clearly related to the relevant concept of murder as are the FBI statistics. It indicates that the efficiency of empirical measures of conviction risk depend heavily on the relevance of the crime statistics used in their construction.

The pattern of results reported in table 5 generally is quite consistent with the pattern observed in the analysis of the FBI data. Weighted regressions, here executed only with $\sqrt{N_S}$ as a weighting factor, yield regression coefficients with lower standard errors than unweighted regressions in both 1940 and 1950. Furthermore, the log-linear regression format is found to be decisively superior to the simple linear format. A Box and Cox test performed on the full observation set in 1940, for example, shows that the approximate logarithmic transformation, unlike

---

[20] For an elaboration of this point see Ehrlich (1975*b*, pp. 212–13). The conceptual superiority of police-gathered data is generally recognized in the criminology literature; see, e.g., Wolfgang (1961, p. 49).

[21] Legal executions are excluded as are, in 1950 only, homicides through legal intervention of police.

TABLE 5

Homicide Supply Functions: 1940 and 1950, GLS Regressions

Eq.	Year	N	"R^2"	Constant	T	P^{*}_{c}	$P^{*}_{e}\|c$	NW	X	W	AGE	URB	EPOS
Results from subsets of states with positive executions:*													
1........	1940	33	.9527	9.036 (2.42)	−0.226 (−2.07)	−0.562 (−4.26)	$\overline{PX4Q}$ −0.341 (−3.54)	0.432 (6.73)	.. ..	0.262 (0.82)	−2.446 (−2.69)	−0.877 (−4.56)	... ...
2........	1950	36	.9555	−2.106 (−0.71)	−0.271 (−2.15)	−0.253 (−1.45)	$\overline{PX5Q}$ −0.196 (−3.21)	0.401 (6.68)	1.617 (2.92)	0.827 (2.14)	0.098 (0.20)	−0.284 (−0.85)	... ...
Results from full-observation set:†													
3........	1940	44	.9526	−326.5 (−3.45)	−0.276 (−2.73)	−0.551 (−4.60)	$\overline{PX4Q}$ −0.333 (−3.56)	0.467 (8.51)	. .	0.483 (1.59)	−2.098 (−2.37)	−0.818 (−4.41)	333.0 (3.57)
4........	1950	46	.9523	−194.8 (−2.95)	−0.154 (−1.36)	−0.211 (−1.52)	$\overline{PX5Q}$ −0.194 (−2.94)	0.335 (5.38)	1.982 (3.51)	0.348 (0.86)	0.303 (0.59)	0.236 (0.82)	194.0 (2.94)

Note.—All eqq. are weighted by $\sqrt{N_S}$.

* All variables measured in natural logarithms.

† All variables except EPOS transformed via $x^* = (x^\lambda - 1/\lambda)$, where x denotes the natural value of a variable and $\lambda = 0.001$

the simple linear transformation, cannot be rejected as an optimal regression form.

4. Murder, Aggravated Assault, and Robbery—Further Analysis of FBI Data

A complete implementation of the analysis of interrelationships among murder, aggravated assault, and robbery developed in Section IA could not be pursued because of the absence of published data on all the specific conditional probabilities and severities of punishment entering equation (3). Introduction in the regression analysis of deterrence variables associated with a specific crime (P^oc_i, T_i) along with corresponding variables associated with interrelated crimes (P^oc_j, $T_j, j \neq i$) generally has been found to yield insignificant estimates of the effects of P^oc_j and T_j on q_i^o, possibly because the latter are not efficient measures of the relevant conditional probabilities or due to a high degree of partial correlation between estimates of probability and severity of imprisonment for different crimes. Only the murder equations reported in table 6, lines 4 and 5, show some evidence for complementarity relations. The bulk of the analysis on interdependencies among these crimes has been conducted through implementation of the restricted model:

$$q_i^{o(\lambda)} = a + b_{i1}T_i^{(\lambda)} + b_{i2}P^oc_i^{(\lambda)} + b_{i3}P^oe|c^{(\lambda)} + \cdots, \qquad (7)$$

where i = assault and robbery, $P^oe|c$ is the conditional probability of execution for murder, and the remaining relevant variables are those introduced in equations (5) or (6). In this model all the relevant probabilities and severities of punishment for related crimes (j) are identified with, or assumed to be proportionally related to, P^oc_i and T_i, and the coefficient b_3 indicates the direction of interdependencies between murder and robbery or assault. In analyses of subsets of executing states the transformation employed in estimating equation (7) is logarithmic. The results are reported in table 6, equations 1–3. (The murder regression is shown for comparison.) Difficulties in the estimation of cross-deterrence effects of probability and severity of punishment notwithstanding the interdependencies among the three crimes examined may still be reflected in systematic correlations among the residual terms associated with each crime. Indeed, analysis of residuals estimated from the single equations reveals sizable *positive* correlations[22] indicating complementary effects. To take account of these correlations in deriving efficient estimates

---

[22] The correlation matrix of residuals associated with murder, robbery, and assault shows that the correlation coefficients between paired crimes range between +.35 and +.5.

TABLE 6

INTERRELATED CRIMES: SELECTED RESULTS FOR 1950 SINGLE EQUATIONS, LOGARITHMIC FORM

Eq.	Dependent Variable	N	"R^2"	Constant	T_i	$P^{o}c_i$	$P^{o}e'c$	NW	X	W	AGE	URB		
					T_{MUR}	P^{o}_{MUR}	$PX5$							
1. . . .	$q^{o}_{(murder)}$[*]	35	.9581	−4.622 3.950 (−1.17)	−0.459 0.128 (−3.59)	−0.687 0.131 (−5.23)	−0.272 0.057 (−4.74)	0.497 0.070 (7.13)	1.200 0.541 (2.22)	1.226 0.463 (2.65)	0.274 0.623 (0.44)	−0.742 0.357 (−2.08)		
					T_{ASS}	P^{o}_{ASS}	$PX5$							
2. . . .	$q^{o}_{(assault)}$	35	.9818	−5.854 5.707 (−1.03)	−0.416 0.209 (−2.00)	−0.429 0.098 (−4.37)	−0.024 0.108 (−0.22)	0.761 0.110 (6.89)	0.175 0.879 (0.20)	1.694 0.701 (2.42)	−0.173 0.889 (−0.19)	−1.160 0.507 (−2.29)		
					T_{ROB}	P^{o}_{ROB}	$PX5$							
3. . . .	$q^{o}_{(robbery)}$	35	.9819	−6.534 5.672 (−1.15)	−0.570 0.219 (−2.60)	−0.887 0.123 (−7.20)	−0.146 0.091 (−1.61)	0.266 0.105 (2.54)	0.932 0.746 (1.25)	1.309 0.772 (1.69)	−0.593 0.868 (−0.68)	0.439 0.519 (0.85)		
					T_{MUR}	P^{o}_{MUR}	$PX5Q$						T_{ASS}	P^{o}_{ASS}
4. . . .	$q^{o}_{(murder)}$	35	.9633	−3.013 (−0.75)	−0.424 (−3.43)	−0.640 (−4.92)	−0.204 (−2.99)	0.466 (6.25)	0.854 (1.35)	1.062 (2.29)	0.093 (0.15)	−0.841 (−2.33)	−0.063 (−0.45)	−0.098 (−1.55)
					T_{MUR}	P^{o}_{MUR}	$PX5Q$						T_{ROB}	P^{o}_{ROB}
5. . . .	$q^{o}_{(murder)}$	35	.9671	−3.777 (−1.01)	−0.398 (−3.20)	−0.670 (−4.86)	−0.224 (−3.98)	0.515 (7.63)	0.822 (1.59)	1.073 (2.11)	0.192 (0.33)	−0.730 (−1.92)	−0.189 (−1.25)	−0.167 (−2.13)

NOTE.—All eqq. are weighted by $\sqrt{N_{REP}}$, reported statistics are $\hat{\beta}$, $S_{\hat{\beta}}$, and $\hat{\beta}/S_{\hat{\beta}}$ in parentheses.

[*] From table 3, line 5.

of the interrelated regressions, both the subsets of executing states and the full observation sets have been estimated simultaneously via Zellner's seemingly unrelated regression method (SUR). The results are reported in table 7.[23]

The regression results are highly consistent with previously reported findings. The main novel finding is that the elasticities of robbery and aggravated assault, as well as of murder, with respect to measures of the conditional risk of execution are negative. Execution risk has the most sizable effect on the murder rate and the least sizable effect on the frequency of assault. The interpretation may be that aggravated assault is more of a "substitute" for murder than is robbery. This pattern of the results is seen more convincingly in the results derived from the application of the SUR method. The standard errors of SUR estimates of regression coefficients, without exception, are lower than the corresponding estimates derived from application of the single-equation least-squares procedure.[24]

5. Pooling of the Cross-Section Samples and Tests of Homogeneity

Not only do the results from 1940 and 1950 independently confirm the basic hypotheses of the model, but the results appear strikingly similar. Combining the two samples permits further testing of the hypothesized effects and allows for systematic examination of the homogeneity of the populations underlying different subsamples. Due to exclusion of the income variables and potential data exigencies, a dummy variable, D_{40}

---

[23] The results reported in tables 6 and 7 are derived from analysis of 1950 data, since in 1940 the application of the SUR analysis was difficult because of the small and variable sample sizes of executing states for which a complete set of data on specific variables is available. The results obtained from application of single-equation least-squares procedures to data from 1940 are qualitatively similar to those reported in connection with the corresponding regressions using 1950 data, but the standard errors of the estimated coefficients associated with $P^{o}e|c$ in assault and robbery regressions were relatively more sizable.

[24] Vital statistics data on the number of homicides occurring in 1950 through legal intervention by police have also permitted an examination of the effect of a measure of the risk of death from police intervention on the frequency of violent crimes. This was achieved by introducing in a regression equation combining murder, robbery, and assault the natural logarithm of the ratio of homicides through police intervention to the estimated total number of murders, robberies, and assaults ($P^{o}d$) as an additional explanatory variable in the context of the model specified in equation (5). The results, based upon the 27 states with both $PX5Q$ and $P^{o}d$ positive, indicate the existence of a negative but weak association between $P^{o}d$ and the rate of violent crimes with the estimated elasticity associated with $P^{o}d$ being -0.225 and its standard error 0.14. The introduction of $P^{o}d$ in the regression analysis did not have any significant effects on the coefficients associated with $PX5Q$ (which is found to be negative and significant) or the other deterrence variables entering the regression equation. Similar results are obtained when $P^{o}d$ is introduced as an additional explanatory variable in regressions relating to the crime of murder alone.

TABLE 7

INTERRELATED CRIMES: 1950, SIMULTANEOUS ESTIMATION SEEMINGLY UNRELATED REGRESSIONS

Eq.	Dependent Variable	Constant	T_i	$P^{o}c_i$	$P^{o}e_i \vert c$	NW	X	W	AGE	URB	EPOS*
Subset of 35 states with $P^{o}e \vert c > 0$; *logarithmic form*†											
			T_{MUR}	P^{o}_{MUR}	$PX5$						
1a..........	q^{o}_{MUR}	−4.378 3.882 (−1.13)	−0.467 0.108 (−4.31)	−0.660 0.113 (−5.85)	−0.273 0.057 (−4.76)	0.494 0.069 (7.13)	1.192 0.522 (2.28)	1.213 0.462 (2.63)	0.255 0.615 (0.42)	−0.748 0.349 (−2.14)	... ...
			T_{ASS}	P^{o}_{ASS}	$PX5$						
1b..........	q^{o}_{ASS}	−6.995 5.662 (−1.24)	−0.375 0.184 (−2.03)	−0.355 0.087 (−4.09)	−0.078 0.103 (−0.76)	0.778 0.109 (7.16)	−0.404 0.843 (0.48)	1.819 0.697 (2.61)	−0.090 0.885 (−0.10)	−1.103 0.502 (−2.19)	... ...
			T_{ROB}	P^{o}_{ROB}	$PX5$						
1c..........	q^{o}_{ROB}	−7.211 5.631 (−1.28)	−0.413 0.195 (−2.12)	−0.780 0.109 (−7.14)	−0.176 0.089 (−1.97)	0.246 0.104 (2.36)	1.183 0.735 (1.61)	1.383 0.755 (1.83)	−0.564 0.868 (−0.65)	0.458 0.512 (0.89)	... ...

Full observation set (44 states); approximate logarithmic form‡

			T_{MUR}	P^{o}_{MUR}	$PX5Q$						
2a..........	q^{o}_{MUR}	−281.85 60.00 (−4.62)	−0.300 0.101 (−2.96)	−0.541 0.100 (−5.43)	−0.278 0.061 (−4.51)	0.431 0.071 (6.04)	1.389 0.551 (2.52)	0.897 0.477 (1.88)	0.710 0.609 (1.17)	−0.416 0.329 (−1.27)	399.75 88.46 (4.52)
			T_{ASS}	P^{o}_{ASS}	$PX5Q$						
2b..........	q^{o}_{ASS}	−107.6 94.66 (−1.13)	−0.393 0.169 (−2.32)	−0.351 0.079 (−4.46)	−0.101 0.094 (−1.07)	0.733 0.096 (7.64)	0.315 0.778 (0.41)	1.543 0.635 (2.43)	0.163 0.789 (0.21)	−0.915 0.429 (−2.13)	145.71 135.95 (1.07)
			T_{ROB}	P^{o}_{ROB}	$PX5Q$						
2c..........	q^{o}_{ROB}	−166.77 81.19 (−2.05)	−0.433 0.175 (−2.48)	−0.831 0.100 (−8.31)	−0.160 0.081 (−1.96)	0.261 0.090 (2.89)	0.922 0.677 (1.36)	1.420 0.660 (2.15)	−0.588 0.755 (−0.78)	0.338 0.426 (0.79)	229.78 117.10 (1.96)

NOTE.—All eqq. weighted by $\sqrt{N_{\mathrm{REP}}}$; reported statistics are $\hat{\beta}$, $S_{\hat{\beta}}$, and $\hat{\beta}/S_{\hat{\beta}}$ in parentheses.

* EPOS is defined here as (2) in executing states and (1) in states where $P^{o}_{e}|_{c} = 0$

† All variables measured in natural logarithms.

‡ All variables (here including EPOS) transformed as follows: $x^{*} = (x^{\lambda} - 1/\lambda)$, where x denotes the natural value of a variable and $\lambda = 0.001$.

(1940 = 1), has been added to the regressions reported in table 8.[25] The results are rather decisive in that they show that the restricted estimation via the pooled samples greatly improves the efficiency of the point estimates of all the deterrence variables. Indeed, the pooled sample results provide, perhaps, the strongest evidence for the deterrence hypothesis. Estimates of the elasticity associated with $P^oe|c$ are nearly the same as corresponding estimates reported in table 3, but their confidence intervals are markedly narrower. Point estimates of comparable deterrent effects are found to be statistically indistinguishable in regressions using the full observation set and the subset of executing states, which is consistent with the results reported in Part 2 of this subsection. Moreover, the differences between the estimated elasticities associated with P^oc and $P^oe|c$ are statistically significant even at the 1 percent level in all the regressions but the southern subset. The analysis also indicates remarkable stability in the basic regression results derived from the 1940 and the 1950 samples. The only apparent difference between the 1940 and 1950 regression estimates reported in table 8 concerns the constant terms of the regressions, as indicated by the effect of the dummy variable.[26] The effect of D_{40} seems to be due to data differences and changes in the relevant income-distribution measures (see n. 25). It is also possible that the general downward trend in murder rates between 1940 and 1950 is partly due to the effect of improved medical technology—a factor which is not accounted for directly in the regression analysis. Application of Chow's test of equality between all other regression coefficients entering equations 1 and 3 in table 8 for 1940 and

---

[25] Because of substantial differences in the definitions of W and X between years, these variables are excluded in the regressions below. Curiously, however, when

$$\begin{pmatrix} X_{40} \\ X_{50} \end{pmatrix} \text{ and } \begin{pmatrix} W_{40} \\ W_{50} \end{pmatrix},$$

along with D_{40}, are introduced into the pooled regressions as nondifferentiated vectors, both variables are found to have the expected positive and statistically significant effects, although the effects of the other variables in equations 2 and 4 of table 8 remain virtually unchanged. Moreover, the standard errors of the coefficients associated with the three deterrence variables and NW are reduced and the inferences concerning tests of homogeneity remain unchanged. Similar findings are obtained in less restricted regressions allowing the coefficients associated with X_{40} and W_{40} to differ from those associated with X_{50} and W_{50}, with the exceptions that the effects of X_{40} and, more notably, D_{40}, become insignificant. Detailed results are omitted here due to lack of space. To achieve greater comparability in variable definitions, the urbanization measure relating to the 1950 sample is defined in all the pooled regressions by the Bureau of the Census's "old," i.e., 1940 definition (URBO_{50}), rather than by the "new" definition used in the preceding 1950 regressions. Remaining potential data differences concern the variable T_{50}, used in both years, and the *UCR* definitions of "urban areas" from which samples of the dependent variables are taken.

[26] It should be pointed out that there is a high degree of competition between D_{40} and AGE in the pooled regressions; the simple correlation between these two variables is about .8. Evidently, the percent of the age group 15–24 in the population of most states was systematically lower in 1950 than in 1940. Hence the sensitivity of the effect of AGE to inclusion of D_{40} in equations 2 and 4 in table 8.

TABLE 8

MURDER SUPPLY FUNCTIONS: POOLED 1940–50 SAMPLE

Eq.	Set	N	"R^2"	Constant	T	P^oc	$P^oe\|c$	NW	AGE	URB	EPOS	D_{40}
							$\overline{PX4Q}$					
1*	Full	87	.9271	−270.8 (−5.183)	−0.304 (−3.599)	−0.616 (−6.828)	−0.270 (−5.206)	0.443 (9.634)	1.252 (5.120)	−0.455 (−3.672)	269.6 (5.212)	... ...
2*	Full	87	.9337	−301.0 (−5.863)	−0.340 (−4.143)	−0.635 (−7.308)	−0.304 (−5.927)	0.460 (10.318)	0.043 (0.087)	−0.562 (−4.491)	303.5 (5.933)	0.362 (2.773)
3†	$P^oe\|c > 0$	67	.9145	−0.939 (−0.716)	−0.346 (−3.552)	−0.654 (−6.173)	−0.269 (−4.898)	0.431 (−7.947)	1.226 (4.421)	−0.454 (−3.199)	... ...	... ...
4†	$P^oe\|c > 0$	67	.9279	3.847 (2.038)	−0.394 (−4.314)	−0.700 (−6.985)	−0.311 (−5.939)	0.453 (8.950)	−0.347 (−0.643)	−0.607 (−4.358)	.. ...	0.469 (3.311)
5*	North (full)	58	.9201	−266.3 (−4.270)	−0.222 (−1.590)	−0.635 (−5.910)	−0.269 (−4.312)	0.551 (8.839)	0.368 (0.541)	−0.887 (−4.682)	268.6 (4.316)	0.328 (1.823)
6*	South (full)	29	.9502	−442.0 (−3.667)	−0.425 (−3.534)	−0.587 (−3.900)	−0.444 (−3.689)	0.456 (4.215)	0.581 (0.487)	−0.485 (−1.823)	442.7 (3.688)	0.171 (0.790)
7†	North ($P^oe\|c > 0$)	39	.9341	4.087 (1.810)	−0.251 (−1.498)	−0.687 (−5.236)	−0.285 (−4.594)	0.602 (8.060)	−0.013 (−0.018)	−1.080 (−4.893)	. ...	0.488 (2.538)
8†	South ($P^oe\|c > 0$)	28	.9447	0.695 (0.166)	−0.426 (−3.530)	−0.584 (−3.895)	−0.442 (−3.686)	0.457 (4.214)	0.589 (0.492)	−0.484 (−1.813)	... ...	0.169 (0.782)
					ΔP^oc	$\Delta P^oe\|c$	ΔNW	ΔAGE	ΔURB			
						$\overline{PX4Q}$						
9‡	$P^oe\|c > 0$ in 1940 and 1950	28	.7504	0.389 (2.597)	−0.642 (−5.919)	−0.256 (−4.682)	0.007 (0.026)	−0.422 (−0.720)	−0.048 (−0.098)	.. ...	... ...	... ...

NOTE.—GLS regressions, selected transformations of linear form, $\hat{\beta}/S_{\hat{\beta}}$ in parentheses Eqq 1–8 weighted by $\sqrt{N_{\mathrm{URB}}}$ For comparability, URB in 1950 is defined as URBO (see n. 25).

* All variables, except EPOS and D_{40}, transformed via $X^* = (X^\lambda - 1/\lambda)$ where X denotes the natural value of a variable and $\lambda = 0.001$.

† All variables measured in natural logarithms

‡ All variables measured as $X^* = \ln(X_{40}) - \ln(X_{50})$, where X denotes the natural value of a variable, and are weighted by $\sqrt{(N_{\mathrm{URB}_{40}} + N_{\mathrm{URBO}_{50}})/2}$.

1950 produces decisive results: the relevant F-statistics are $F_{7,71} = 1.50$ and $F_{6,53} = 1.804$, respectively. Clearly, then, the hypothesis of no significant differences cannot be rejected at the 5 percent level of significance. Similar results are obtained through regression analyses based on alternative weighting factors and alternative execution risk measures. Results concerning tests of equality between coefficients associated with deterrence variables only are even sharper.

The homogeneity of populations pooled in the regression analysis performed in this study has also been tested across the North and South regions.[27] The results are found to be highly consistent. Comprehensive tests of equality between all the regression coefficients associated with northern and southern samples indicate that the hypothesis of equality cannot be rejected at the 5 percent level of significance. These results are corroborated across samples from 1940 and 1950 as well as the pooled sample (see, for example, eqq. 5 and 6 in table 8). The apparent homogeneity indicates that the regression results reported are not an artifact of *geographically* specific factors unaccounted for in the regression analysis.

In principle, the sharp set of results summarized in equations 1–8 in table 8 may be due to the fact that the 1940 and the 1950 samples are "identical" or that the vectors of variables are strictly proportional, in which case an apparent improvement of regression results is expected on spurious grounds. Examination of the distributions of the key deterrence variables reveals, however, considerable changes in the relative distributions. A stronger confirmation of the technical independence of the samples is provided through regressions based on differences in the values of the relevant variables within the same states between 1940 and 1950. The results, illustrated in equation (9) for the set of consistently executing states, are highly compatible with all previously reported results, as are other results not included here because of lack of space. The effects of deterrence variables are found to be quite similar to those estimated in the cross-sectional analysis, but the effects of demographic variables are considerably less pronounced, presumably because of the relatively smaller changes in the relative distributions of these variables. The "disappearance" of the effect of *NW* in these regressions is compatible, however, with the results from the time-series investigation (Ehrlich 1975*a*, p. 412). It suggests that the significant effect of this variable in the cross-sectional investigations may largely reflect the effects associated with relatively low earning opportunities for nonwhites that are not captured by *W* and *X*. Over time, these opportunities have improved. The main contribution of the regression analysis of murder changes is its compatibility with the theorized predictions concerning the signs and relative magnitudes of the elasticities of the murder rate with respect to $P^{o}c$ and $P^{o}e|c$ and its consistency with the cross-sectional results on deterrence. It clearly suggests that the strong results derived from the pooled samples are not artifacts of "sample duplications."

---

[27] States designated South in the 1940 and 1950 samples (constrained by data availability) include Alabama, Arkansas, Delaware, Florida, Kentucky, Louisiana, Maryland, Mississippi, North Carolina, Oklahoma, South Carolina, Tennessee, Texas, Virginia, and West Virginia. The same designation is applied in 1940 and 1950 with the exception that Mississippi is not included in the 1940 sample. Note that the introduction of EPOS in the southern subset (eq. 6 in table 8) is due to the fact that for only one state, Delaware, $PX4Q$ is zero in 1950. In related regressions using $PX5$ as the conditional risk of execution measure, all southern states, including Delaware, are classified as executing states. The results, nevertheless, are highly consistent with those reported in equations 6 and 8. The fact that EPOS has the usual positive and statistically significant effect even in equation 6 therefore underscores the argument that the appearance of significant differences between constant terms in "executing" and "nonexecuting" states is largely due to the assignment of inappropriate values to measures of $Pe|c$ in such states. Conceivably, the constant term in the regression equation associated with Delaware is estimated to be lower than in other states because the true level of execution risk in Delaware had been positive, not zero as equation 6 implies.

III. Executing and Nonexecuting States

The murder rate has been shown to depend, in part, on the risks of conviction and execution and on the severity of imprisonment for murder, each of which, in turn, is subject to the control of public authorities. As is pointed out in Appendix B1, however, the magnitudes of these variables may be the result, rather than the cause, of the relevant crime rates. In particular, the tendency of state courts, juries, and other relevant authorities to retain or impose the death penalty is expected to be responsive to the perceived social costs of capital crimes. The purpose of this section is to compare some crime and punishment data across executing and nonexecuting states in light of theoretical expectations.

In the analysis of optimal defense against crime, law enforcement agencies are assumed to behave in accordance with the expectation that certainty and severity of imprisonment and the risk of execution have restraining (i.e., deterrent and incapacitating) effects on the frequency of crime. The agencies are assumed to set optimal penalties and produce optimal values of the probability of conviction and the conditional probabilities of specific penalties so as to minimize the net per capita social losses from crime, including the costs of all law enforcement activities. The components of the social loss function are identified in equation (8) below:[28]

$$L = D(q) + C(q, Pc) + \gamma_1 Pc\, Pe|c\, qd + \gamma_2 Pc(1 - Pe|c)m. \qquad (8)$$

---

[28] In a more general formulation of the model, one may wish to define q as a vector of interrelated crimes such as murder, robbery, and assault. Similarly, Pc, m, and $Pe|c$ could be defined as vectors of corresponding deterrence variables. Such generalization would complicate the derivations of the basic behavioral implications discussed but would not change their basic thrust. Behavioral implications following from the formulation adopted in equation (8) are formally developed in the mathematical Appendix to this paper, omitted here for lack of space. One important implication of that model, which is compatible with the empirical findings reported in Section II, is that the elasticity of the murder rate with respect to Pc exceeds the corresponding elasticities with respect to both $Pe|c$ and the cost of imprisonment m, or $\epsilon_{Pe|c} < \epsilon_{Pc} > \epsilon_m$.

In equation (8) q is the crime rate, $D(q)$ designates the net direct social damage from crime, $C(q, Pc)$ represents the costs of apprehending and convicting offenders, and the last two terms represent the social costs associated with executing and incarcerating convicted persons, respectively. The terms d (treated as a parameter at a point in time) and m measure the discounted value of the private costs of execution and imprisonment per convict, and γ_1 and γ_2 are positive coefficients "translating" private into social (deadweight) costs and augmenting the relevant public expenditures in administering penalties. Both $D(q)$ and $C(q, Pc)$ are assumed to be convex functions.

In terms of the social loss function introduced in equation (8), optimal (cost-minimizing) values of the risk or status of execution as well as the other deterrence variables must equate the marginal cost of deterrence with the marginal revenue from the expected decrease in crime levels. An important implication of the optimality conditions is that the optimal risk of execution could be considerably below unity even if on the margin the death penalty has a differential restraining impact on murder over and above that of imprisonment because executions entail a larger deadweight loss on society (the difference between $\gamma_1 d$ and $\gamma_2 m$ is assumed positive). Variations in the enforcement level of capital punishment, including its abolition, as well as changes in the magnitudes of other deterrence variables generally can be explained as a consequence of shifts in three basic sets of parameters: the level of the marginal harm from murder and related crimes, the social cost of capital punishment relative to other sanctions, and the technology of producing means of deterrence and incapacitation. The first two sets of factors appear to be especially relevant for analysis of data relating to the classes of executing and nonexecuting states.

A central behavioral implication of the model is that the optimal magnitudes of all three deterrence variables generally must rise as a result of an increase in the marginal social damages from crime due to exogenous factors. Since stability conditions for optimal law enforcement are satisfied if, among other related conditions, the marginal damages from crime are monotonically related to the rate of crime in the population,[29] an increase in the crime rate due to exogenous factors would require that the optimal values of deterrence variables be increased. A decrease in the crime rate then would have the reverse effects. While the theory does not predict that each of the three deterrence variables would be reduced as a consequence, it suggests that both certainty of conviction and severity of punishment would tend to decline if "punishment" were defined as a single-choice variable. More generally, $Pe|c$ and m always are expected to move in the same direction if γ_1 and γ_2 remain constant. In particular, it follows that those states where murder rates have traditionally been exceptionally low due to factors exogenous to the law enforcement system (such as demographic and general economic variables) are more likely to minimize the application of, or abolish, the death penalty either legally or in practice. These states also are expected to impose relatively short imprisonment terms and possibly to have relatively low conviction risks.

---

[29] This assumption can be justified partly on the grounds that the resource costs expended by offenders in the course of their criminal activity rise with the frequency of crime committed (especially by new "entrants") and because the marginal costs of apprehending and convicting offenders are expected to be an increasing function of the frequency of offenses committed (both C_{qq} and C_{qPc} are positive). Indeed, it can be shown that if both the production function of law enforcement activity implicit in $C(q, Pc)$ and the supply-of-offenses function are of the Cobb-Douglas variety, then an exogenous increase in q will necessarily lead to an increase in the magnitudes of severity and certainty of punishment even if $D_{qq} = 0$. Of course, the marginal social valuation of the human and nonhuman assets destroyed through crime must eventually rise as their "stocks" are depleted with more crimes committed at a point in time.

The theory allows for corner solutions in the risk of execution across states to occur not only, however, as a consequence of autonomous changes in the crime rate, but also as a result of changes in factors determining the "social cost" of execution. For example, changes in political factors strengthening the influence of groups interested in greater emphasis on rehabilitation or reformation of offenders, federal judicial or administrative rulings, or increased social aversion toward harsh punishments would either raise the magnitude of the coefficient γ_1 in equation (8) or might lead directly to a decrease in the magnitude of $Pe|c$. Under such circumstances, if a corner solution occurs so that execution is abolished de facto, the theory predicts that either the probability of conviction or the severity of punishment by imprisonment or both will be increased. The intuitive reason is that to the extent that cessation of execution is not motivated by a low level of crime, relinquishing the practice of execution must be "compensated" through greater reliance on alternative instruments of law enforcement in order to bring about an optimal level of deterrence and incapacitation.

Both predictions appear to be compatible with data on the mean value of crime and deterrence variables across various subsets of executing and nonexecuting states in 1940, 1950, and 1960. As table 9 demonstrates, in all 3 years analyzed, the murder rate is found to be lowest in states which abolish the death penalty in law, somewhat higher in states which persistently refrain from enforcing the death penalty, and is even higher in states which only intermittently refrain from executing convicts. In addition, the murder rate is always higher in the corresponding subsets of executing states which complement the three subsets of legally or practically nonexecuting states discussed above. Moreover, subsets of abolition states having exceptionally low murder rates not only choose to refrain from using the death penalty, they also tend to set somewhat lower

TABLE 9

MEANS OF ESTIMATED CRIME RATES, PROBABILITIES OF CONVICTION, AND IMPRISONMENT TERMS IN EXECUTING AND NONEXECUTING STATES, 1940–60

SUBSETS COMPARED BY YEAR AND MEAN*	EXECUTION STATUS IN t		EXECUTION STATUS IN 5 YEARS		LEGAL STATUS OF DEATH PENALTY	
	$E_t > 0$ (1)	$E_t = 0$ (2)	$\sum_{j=0}^{4} E_{t-j} > 0$ (3)	$\sum_{j=0}^{4} E_{t-j} = 0$ (4)	Legal (5)	Illegal (6)
1940:						
q^* (murder)	9.9069	2.0947	8.5583	1.5833	7.6048	1.2833
P^*_c (murder)†	0.2074	0.3438	0.2338	0.3523	0.2597	0.2739
T_{50} (murder)	119.759	158.105	130.222	149.083	135.143	133.500
q^* (assault)	87.0034	17.6421	75.0028	13.1833	66.0595	13.9667
q^* (robbery)	60.0414	29.0737	55.6083	24.3083	51.0309	25.0500
$PX5$‡	. .	..	0.1552	...	...	...
No. of states	29	19	36	12	42	6
1950:						
q^* (murder)	8.5291	2.3752	6.3541	1.8591	5.8679	1.5167
P^*_c (murder)§	0.2324	0.3106	0.2595	0.3183	0.2735	0.2514
T (murder)	117.696	150.800	135.676	132.445	135.143	133.500
q^* (assault)	105.004	27.0960	76.0486	25.3363	69.8405	26.5333
q^* (robbery)	49.6522	31.1000	43.1378	29.4000	42.3809	23.2500
$PX5$ǁ	...	...	0.0809	...	...	...
No. of states	23	25	37	11	42	6

1960:**						
q^{o} (murder)	7.1050	3.3857	6.2492	2.2750	5.3397	2.3857
P^{o}_{c} (murder)††	0.4548	0.4830	0.4542	0.5034	0.4776	0.4330
T (murder)††	134.890	169.104	131.361	199.642	144.277	213.214
q^{o} (assault)	71.9200	40.7849	69.8000	21.7875	58.6103	25.4428
q^{o} (robbery)	40.1300	33.2727	43.4018	21.7875	38.1927	24.4857
$PX5$‡‡	..	...	0.0195	...	...	..
No. of states	20	28	32	16	41	7

---

* Simple means of the state statistics; number of states compared dictated by the availability of relevant data.

† Cols. 1 and 3 exclude Georgia, Mississippi, and Nevada, cols. 2 and 4 exclude Maine and Vermont, col 5 excludes Georgia, Mississippi, Nevada, and Vermont; col 6 excludes Maine.

‡ Col 3 estimate excludes Georgia and Mississippi

§ Col. 1 excludes Georgia, col. 2 excludes North Dakota, Michigan, and Vermont; cols 3 and 5 exclude Georgia and Vermont, cols. 4 and 6 exclude Michigan and North Dakota.

‖ Col 3 estimate excludes Georgia.

** Data for 1960 exclude Alaska and Hawaii. The data for 1960 are not comparable to the 1940 and 1950 data due to the following differences (1) In 1960 q^{o} is based on total reported crimes in each state, whereas q^{o} in 1940 and 1950 is based on a state sample of urban areas (2) Commitments to state prisons in 1960 are convicts of homicide whereas in 1940 and 1950 commitments are for murder convicts. Commitments in 1960 include only convicts with definite sentences. To estimate convictions, E_{60} is added to commitments. Also, P^{o}_{c} in 1960 is based on total reported crimes, Q^{o}, whereas in 1940 and 1950 P^{o}_{c} is based on $Q^{o} = q^{o} \times N_{s}$ (3) T in 1960 is average time served whereas median time served measures T in 1950.

†† Cols. 2, 3 and 5 exclude New Jersey.

‡‡ Col. 3 estimate excludes New Jersey

values of conviction risks and imprisonment terms.[30] Indeed, there is a strong indication—apart from the regression results in Section II—that the relatively low magnitudes of the execution risk variables in these subsets are the result rather than the cause of low murder rates. As table 9 shows, the subsets of persistently nonexecuting states are characterized by much lower frequencies of assault and robbery relative to the corresponding executing or retentionist states in all 3 years considered. Given the regression results, the differences between the crime levels in the two classes of states may be ascribed, in part, to the substantial differences in demographic variables, such as the percentage of nonwhites, and possibly to other "missing" or misspecified factors.[31] To some extent, the relatively low frequencies of assault and robbery may even be contributing causes for the legal or practical abolition of the death penalty, because murder is often a by-product of assault and robbery and the enforcement of capital punishment may have restraining effects on these crimes as well (see Section II). Thus the incentive to enforce the death penalty might be monotonically related to the frequencies of assault and robbery as well as of murder. Again, there is some evidence that at least severity of imprisonment for assault and robbery tends to be lower in abolitionist than in retentionist states.

At the same time, table 9 reveals that in the set of states which refrain from enforcing the death penalty in a given year, but not necessarily over a greater period, and in which murder rates are found to be higher than in the relatively smaller subset of persistently nonexecuting states, both probability and severity of imprisonment for murder appear to be higher than in the complementary set of executing states. This is the case in all 3 years considered. Furthermore, in 1960 even the set of persistently nonexecuting states, considerably larger in 1960 than in 1950, is characterized by higher values of probability and severity of imprisonment for murder relative to the complementary set of executing states.[32] This tendency for greater reliance on certainty and severity of imprisonment in specific subsets of nonexecuting states, especially in more recent years, is compatible with the theoretical expectation that reductions in execution risk prompted by an increase in the social cost of execution and related exogenous factors would result in compensatory changes in certainty and severity of alternative punishments. The existence of such a tendency is not inconsistent with the hypothesis that capital punishment and other penalties do exert a restraining effect on the murder rate. It is at least compatible with the assumption that law enforcement bodies behave in accordance with that hypothesis.[33]

---

[30] Due to the small number of states included in the subsets of states where execution is illegal, the means of the state statistics based on single-years data must be viewed with caution. (See also n. 32 below.)

[31] The simple mean levels of *NW* in executing (cols. 3 and 5 in table 9) and nonexecuting (cols 4 and 6 in table 9) subsets are 12.3 and 11.0 as against 2.0 and 2.2 in 1950, and 12.9 and 11.4 as against 2.2 and 1.6 in 1940, respectively. The corresponding mean levels of A_{25-34} also are slightly lower in nonexecuting states in both years. It also seems that abolitionist states tend to be more urbanized than retentionist states in both 1940 and 1950. Other variables examined in the regression analysis do not differ markedly between the two subsets.

[32] Average and median imprisonment terms served for murder in nonexecuting states may be higher than those in executing states simply because murder convicts liable for capital punishment receive long prison sentences in the former class. That this technical consideration may not provide the main explanation for the observed differences in actual imprisonment terms between executing and nonexecuting states is apparent from the fact that states where the death penalty was illegal in 1950 did not have a higher median imprisonment term than states which still retained the death penalty in that year.

IV. Summary and Conclusion

Contrary to strong inferences of many death-penalty researchers who interpreted their evidence as a categorical denial of the deterrence hypothesis, this study as well as a related analysis based on time-series data develop evidence not inconsistent with that hypothesis. The case for the deterrence theory is based, in part, on a critical evaluation of earlier works (see n. 1). In the main, the case is based on new empirical findings and their compatibility with a set of predictions derived from the theory.

Before summarizing the main findings of the empirical investigation, some of its inherent limitations must be stressed. The present analysis does not account systematically for potential simultaneity relations between crime rates and deterrence variables—especially the risks of execution and conviction. Although the simple remedies to the problem invoked here may lead to an understatement of the true effect of execution risk, the net direction of the bias cannot be ascertained with confidence because the estimates of effects associated with other deterrence variables may not be statistically consistent. There are many imperfections associated with measures of the probabilities of conviction and execution and the severity of imprisonment as viewed by potential offenders. The regression equations do not include estimates of the extent and quality of medical services available across states, and the interdependencies among murder, robbery, and aggravated assault are dealt with only in a partial manner. While the present investigation deals with the choice of an optimal regression format, it does so only in the context of restricted families of "power transformations." All the statistical tests conducted are conditional upon the validity of the assumptions concerning disturbance terms.

---

[33] Although this analysis of optimal law enforcement against murder has been applied here in explanation of changes in execution status across different states, it also can be applied in explanation of trends in the enforcement and status of capital punishment in connection with murder and other "capital crimes" over time. An attempt along these lines is included in a related unpublished paper where the social loss from offenses is generalized to incorporate a criterion of "equity" in addition to that of efficiency.

At the same time the consistency and stability in the results obtained upon application of efficient estimation procedures, given the sample limitations, and their consistency with specific theoretical predictions and previous findings seem remarkable. Indeed, it is noteworthy that all the deterrence variables examined in this analysis yield the expected results in connection with murder and other crimes, and that the coefficients associated with explanatory variables other than constant terms appear statistically indistinguishable across different samples as well as across subsets of executing and nonexecuting states. In particular, the subsets based on data from executing states indicate that when enforced the death penalty exerts a restraining effect on the frequency of murder and possibly robbery as well. The investigation shows the existence of distinct execution and imprisonment effects: both $P^o{}_{e|c}$ and T appear to exert a deterrent effect on murder. Indeed, because the effect of imprisonment is accounted for explicitly in the analysis, this investigation, more directly than the earlier time-series study, indicates that capital punishment has a differential deterrent effect over and above the actually enforced imprisonment terms. The effects of execution risk and other deterrence variables are largely unaffected by the introduction of additional explanatory variables, including the "prices" associated with related crimes and other indicators of relative rewards. Even introduction of risk of death through police intervention (see n. 24) does not affect the regression results. Moreover, as predicted by the theory, the estimated elasticities of murder rates with respect to measures of conviction risk are significantly larger than the estimated elasticities associated with measures of the conditional risk of execution. Also, that the estimated elasticity of the murder rate with respect to conviction risk exceeds the corresponding elasticity with respect to imprisonment severity is not inconsistent with the analysis of optimal public defense against murder (see n. 28). The joint estimation of "supply functions" associated with murder, robbery, and aggravated assault, as well as the pooling of the 1940 and the 1950 samples, narrow the confidence intervals associated with point estimates. Furthermore, the analysis of changes in murder rates within states between 1940 and 1950 produces results which corroborate the general cross-sectional findings.

More important, the basic set of results derived in this cross-sectional investigation is in essence the same as that derived from analysis of time-series data on murder, although the results have been obtained through different estimation procedures and although the time-series analysis did not account for possible variations in imprisonment severity and other relevant variables. The corroborating independent evidence from the 1940 and the 1950 cross sections does not support the proposition that the time-series results on deterrence are merely artifacts of the 1960s or the result of "missing factors" such as the Vietnam War or the proliferation of handguns. Despite statistical differences, point estimates of the elasticity of the murder rate with respect to measures of unconditional probabilities of conviction in the time-series and cross-section analyses lie within each other's confidence intervals (compare, e.g., the results in Ehrlich [1973, table 7] and the results reported above in tables 3, 4, and 8). The larger point estimates of the effect of the unconditional risk of execution on the extent of the murder rate in the cross-sectional analysis may be the result of the fact that the data permit the estimation of such an effect within the set of states with positive executions, whereas the aggregate data used in the time-series analysis have not allowed for a separate estimation of deterrence effects within strictly execution states.[34] Tentative estimates of the average trade-offs between executions and murders computed via the method employed in the time-series study (Ehrlich 1975*a*, section IIIB) on the basis of the GLS results reported in equations 1–4 in table 8, for example, range between 20 and 24 murders for one execution. It is important to stress that these estimates are based upon regression coefficients derived through the GLS estimation procedure which does not insure consistent estimates. They also are conditional upon specific data limitations.[35]

One of the conspicuous results of the cross-sectional investigation is the strong statistical association between crime rates, especially murder and assault, and the percent of nonwhites in the population. However, the NW effect may, in part, reflect the roles of other demographic variables not fully analyzed in this paper. Its magnitude is reduced when A_{25-34} replaces the usual AGE (15–24) variable. The NW effect in the cross section may largely reflect the effect of low earnings opportunities, specific to nonwhites, which are not accounted for directly in this analysis. The relevance of the latter factor is indicated by the fact that both my time-series analysis and the examination of murder changes between 1940 and 1950 fail to uncover any significant association between NW and murder over time.

---

[34] See the discussion in Ehrlich (1975*a*, p. 408). It is possible that the cross-sectional elasticities are biased upward relative to the time-series results due to "spillover effects" (see Ehrlich 1974, n. 25) that result from interstate movements of offenders in response to interstate variations in deterrence variables. However, the results show that the estimated elasticities associated with conviction risk are much more similar across the time-series and the cross-section analysis. More generally, it seems unlikely that "spillover effects" would be empirically important in connection with crimes against the person. It should be pointed out that the estimated elasticity of the murder rate with respect to $Pe|c$, in principle, includes both deterrent and incapacitation effects. An estimate of the latter derived through a procedure similar to that used in Ehrlich (1975*a*, p. 413) indicates that the magnitude of the incapacitation effect is at most 0.08, much lower than the estimated elasticities.

[35] It should be emphasized that in 1940 and 1950 there do not exist estimates of the total number of murders within individual states so that even tentative tradeoffs cannot be calculated without additional assumptions. The numbers mentioned above are implied by the relevant elasticities associated with measures of execution risk on the assumption that urban and rural murder rates were identical in the states included in the general sample underlying table 8.

Another intriguing issue concerns interdependencies between levels of crime rates and deterrence measures in executing and nonexecuting states. Consistent with analysis of optimal social defense against crime, the persistently nonexecuting states face substantially lower risks of victimization through murder and related crimes than do the persistently executing states. The low crime levels in the former class of states, in turn, can be partly attributed to factors whose influence is reflected by the percentage of nonwhites in the population and related demographic variables. Some of the apparent differences in crime levels across the two classes of states, as revealed through the regression model, also may result from understatement of the true (perceived) risks of execution in states classified as "zero" execution risk states. A fuller analysis of the differences in crime levels across these classes and of the variations in execution policy across states and over time must await future research.

Hopefully, the main contribution of this research lies in the suggestion that the basic economic and econometric frameworks used by economists to explain behavior in the marketplace can also be applied in explaining criminal and perhaps some other behavior traditionally labeled as "deviant." The regularities uncovered in connection with movements in crime rates and law enforcement activities pose a challenge for future research. The economic approach might prove useful in analyzing recidivism by offenders, the impact of legitimate employment and training opportunities, apparent racial and sex differences in participation in specific crimes, international variations in crime rates, and public and private protection against crime. As to the policy implications emanating from the new findings, however, basic theoretical and methodological questions remain unanswered because of the complexity of the range of choices available and the difficulties associated with quantifying basic parameters affecting the social costs from offenses and other targets of social policy (see n. 33). Estimates of the restraining effect of the risk of execution in terms of the implied trade-offs between lives sacrificed and lives saved, though apparently not negligible, must be viewed as tentative and imprecise also because they may not represent the relevant *marginal* trade-offs. But even if the estimates were accurate, they simply might indicate the extent of the full social costs associated with sacrifices of the lives of convicted persons for the lives of potential victims, assuming that the actual policy toward capital punishment has reflected a local social optimum. Perhaps the main policy implication of the present study lies in the support it lends to the general deterrence hypothesis. Indeed, if punishment did not deter offenders, then incapacitation and reformation of those apprehended and convicted would become the sole means of crime control. Not only would such means be of limited usefulness because of the small fraction of potential offenders who would be incapacitated or reformed at a point in time given the relevant costs, but incapacitating penalties are intrinsically inefficient modes of taxing illegitimate behavior from a social point of view. In contrast, if offenders do respond to incentives, punishment and law enforcement become more expedient means of crime control independently of any incapacitating or reformatory benefits of sanctions. Indeed, incapacitation and related "cruel" punishment, in principle, can be replaced by more "humane" penalties involving smaller deadweight losses. Moreover, crime control could be achieved through an optimal balance between both negative and positive incentives.

Appendix A

An Analysis of Transformations

Assuming that the correct functional form of the supply-of-offenses function belongs to the following single-parameter class of "power transformations" involving all the variables of relevance, one may write that function

$$q_i^{(\lambda)} = \alpha + \beta_1 Pc_i^{(\lambda)} + \beta_2 Pe|c_i^{(\lambda)} + \cdots + u_i, \tag{A1}$$

where the operator (λ) associated with any variable X_i is defined by $(X_i^\lambda - 1)/\lambda$. The $\lambda = 1$ defines the simple linear transformation with all variables represented by their natural values, and $\lambda = 0$ defines the logarithmic-linear transformation. A search for an optimal transformation can be pursued via direct application of the Box and Cox analysis of transformations (1964).

In applying the Box and Cox analysis it is assumed that for some unknown λ that $u_i \sim N(0, \sigma^2)$ and that the variables on the right-hand side of (A1) are uncorrelated with the error term. The determination of an optimal specification is then achieved by first calculating the maximized log likelihood function (excluding constant terms) for alternative values of λ:

$$L_{\max}(\lambda) = -\frac{n}{2}\ln \hat{\sigma}^2(\lambda) + (\lambda - 1)\sum_{i=1}^{n} \ln q_i. \tag{A2}$$

After determining the maximizing value $\hat{\lambda}$, alternative values of λ can be rejected, using large sample maximum-likelihood theory, by testing whether they belong outside of an approximate $100(1 - \alpha)$ percent confidence interval obtained from:

$$2[L_{\max}(\hat{\lambda}) - L_{\max}(\lambda)] < \chi_1^2(\alpha), \tag{A3}$$

where $\chi_1^2(\alpha)$ denotes the χ^2 statistic with 1 df. It can easily be shown through an application of Box and Cox's analysis that in the case of weighted regressions, equation (A2) becomes:

$$L_{\max}(\lambda) = -\frac{n}{2}\ln \hat{\sigma}'^2(\lambda) + (\lambda - 1)\sum_{i=1}^{n} \ln q_i + \frac{1}{2}\sum_{i=1}^{n} \ln N_i, \tag{A4}$$

where $\hat{\sigma}'^2(\lambda)$ denotes the maximum-likelihood estimate of residual variance calculated from the weighted regressions and the $N_i^{1/2}$ denote the weights. Since tests for homoscedasticity reported in Appendix B show that the error terms in equation (A1) are not homoscedastic, and the analysis of residuals suggests optimal weights for obtaining GLS estimates of equation (A1), both equations (A2) and (A4) are utilized in the pursuant empirical investigation to calculate the relevant log-likelihood functions.

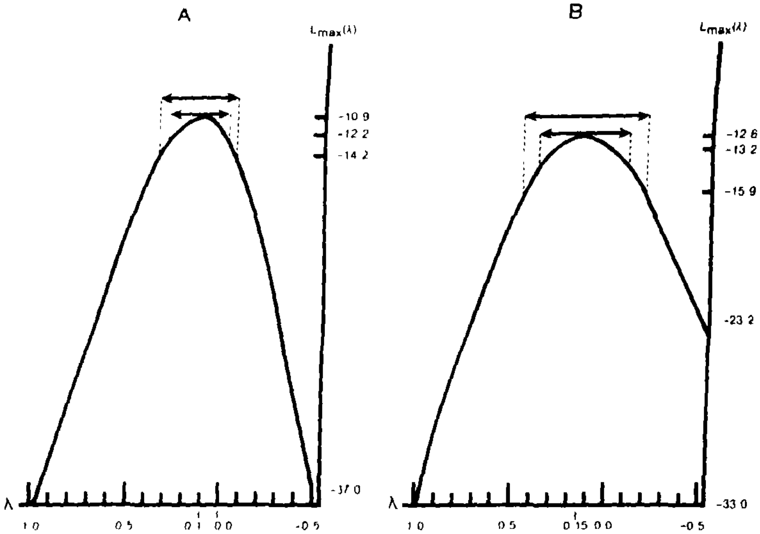


FIG. 1.—Maximized log-likelihood functions for Box and Cox tests

These tests for optimal transformations have been repeatedly applied against data relating to both subsets of states with positive executions and to the full samples of states in both 1940 and 1950. The regression models used in the computation of equations (A2) and (A4) are those underlying equations (5) and (6) in the text, and the tests were conducted with alternative weighting factors and with different estimates of execution risk. A sample of results for selected transformations is reported in table 4, parts A and B. In each table detailed regression results are reported for transformations by $\lambda = 1$ (the simple linear forms), $\lambda = 0.001$ (the approximate log-linear transformation used in connection with the full-observation set), and by $\hat{\lambda}$ (the maximizing value of $L_{max}[\lambda]$). Figure 1 depicts the maximized log-likelihood functions derived from an analysis of one subset of executing states and one full-observation set and indicates the approximate 95 percent and 99 percent confidence intervals associated with $\hat{\lambda}$.

The implications of the Box and Cox tests are rather decisive. Without exception, the tests conducted show that the simple linear form must be rejected as an optimal transformation. The value of L_{max} $(\lambda = 1)$ lies far outside the 99 percent confidence interval around $\hat{\lambda}$. In contrast, in all of the tests performed, the log-likelihood function evaluated at $\lambda = 0$ (in executing subsets) or $\lambda = 0.001$ lies within the relevant 95 percent or the 99 percent confidence intervals.

Furthermore, the same unequivocal results are found in tests of transformations related to regression equations where the pair $(P^oc, P^oe|c)$ is replaced by either (P^oc, P^oe) or by (P^om, P^oe)—the basic "prices" of the model (see the discussion in Sections IA and IC). The regression results derived via the latter specification also are found to be consistent with the theoretical expectations. The effects of both Pc and Pe or Pm and Pe are negative and statistically significant. Moreover, the relative magnitudes of the various deterrent effects estimated via the alternative models are found to be the same as those expected from analysis of equation (1).

The preceding analysis indicates the relative superiority of the logarithmic over the simple linear functional form within the class of transformations (A1) which includes both forms as special cases. In view of the discussion in Section IC and the evidence developed in this Appendix, it also is of interest to evaluate the logarithmic transformation within an alternative class of transformations that includes a semilogarithmic functional form as a special case. That class is given by

$$\ln q = \alpha + \beta_1 Pc^{(\lambda)} + \beta_2 Pe|c^{(\lambda)} + \cdots + u. \tag{A5}$$

Here $\lambda = 1$ defines the conventional semilogarithmic specifications and $\lambda = 0$ defines the logarithmic form. By invoking the assumptions made in the preceding discussion, an analysis of transformations can be pursued through a similar application of the Box and Cox approach. The maximized log likelihood, for fixed λ, is here, except for constant terms

$$L_{\max}(\lambda) = -\frac{n}{2} \ln \hat{\sigma}'^2(\lambda). \tag{A6}$$

In the tests conducted in connection with equation (A5), the value of λ that maximizes equation (6), $\hat{\lambda}$, is much closer to $\lambda = 0$ than to $\lambda = 1$. More important, the 99 percent confidence interval around $\hat{\lambda}$ always includes $\lambda = 0$, whereas $\lambda = 1$ is always far outside that interval and must therefore be rejected as an optimal transformation.

The general conclusion emerging from these analyses is that under the assumptions concerning the distribution of the disturbance terms, the logarithmic transformation cannot be rejected as an optimal transformation within the classes of "power transformations" considered in this analysis.

Appendix B

The Estimation Procedure and Variables Used

1. *The Estimation Procedure*

The implicit assumption underlying the estimation procedure is that the relative variation in exogenous factors controlling demand-for-enforcement and production-of-deterrence functions is sufficiently greater than that of missing exogenous or stochastic factors controlling supply relationships so that the latter can be "traced out" effectively. Previous research indicates that this assumption is not at odds with findings concerning the effect of conviction risk and imprisonment severity: the relevant estimates derived via OLS and TSLS techniques have been compatible.[36] It should be emphasized that, due to data exigencies, the measure used in this investigation to estimate the length of imprisonment anticipated by potential offenders in 1940 is the actual time served by prisoners released in 1951 rather than in 1940 (T_{50}). Neither the 1940 nor the 1951 measure should necessarily give rise to an important simultaneous-equation bias (see n. 36). But the use of T_{50} as an instrument in the 1940 regressions should practically eliminate biases due to a potentially limiting impact of prison "crowding" on the opportunities to execute long imprisonment sentences in states which experienced unanticipated increases in crime rates in 1940. In contrast, use of T_{50} in the 1940 regressions, in principle, might generate biases due to a potential tendency to mete out relatively long prison sentences in states that experienced relative increases in crime rates in 1940 and because the actual T_{50} figures should be inaccurate measures of the anticipated terms in 1940. The fact that the estimated coefficients of T_{50} as well as of other deterrence variables in the 1940 and the 1950 regressions are found to be statistically indistinguishable (see Section IIB5) lends support to the particular assumption that the instrumental measures of imprisonment terms may be considered "exogenous" or predetermined variables.

---

[36] Compare the OLS and TSLS estimates of supply-of-offenses functions in Ehrlich (1970, 1974). Estimates of actual time served in state prisons by prisoners released in sample years have been treated as an exogenous variable even in these earlier analyses partly because of the apparent stability in the distribution of T relative to that of crime rates over periods preceding specific sample years and partly because this variable may in large measure be a true "predetermined" variable in view of the time lag between the date in which offenders are sentenced and their parole eligibility determined and the time of their actual release from prison. Even in the time-series analysis of murder (Ehrlich, 1975*a*), where no account could be taken of changes in imprisonment severity, the results derived through the simultaneous-equation technique have been found to be quite similar to (unpublished) OLS estimates of the supply-of-offenses functions.

A special problem arises in connection with nonexecuting states. In the first place, this class of states, which is at a "corner solution" with respect to use of the death penalty, in practice is found to be an outlier insofar as the rate of crime is concerned. Not only is the murder rate exceptionally low in this class of states, but the rates of robbery and assault there are exceptionally low as well (see table 9). Analysis of optimal defense against murder suggests that declines in the supply of offenses due to exogenous factors would decrease the desire to impose severe sanctions on offenders. Thus, if there were significant shifts in supply functions across subsets of executing and nonexecuting states due to either stochastic factors or to variables not accounted for in the regression analysis, such as medical technology and the degree of social interaction, then these shifts could trace out "demand" rather than "supply" curves (the curves in question relating q to inverse functions of the relevant deterrence variables). Second, and more important, there is a difficulty in assigning quantitative values to the level of execution risk in nonexecuting states. Despite the fact that all states in this class by definition refrain from execution over a substantial period of time, the perceived risk of execution given conviction in these states need not be zero or even identical because of differences in conviction *levels* across states and because of varying expectations concerning future trends. The problem becomes considerably more severe in regressions based on an approximate logarithmic specification: a transformation of $Pe|c = 0$ by the Box and Cox technique with λ set at an arbitrarily small value creates a very large difference between the values of $Pe|c$ in executing and nonexecuting states. Such transformation of the partly null vector of $Pe|c$ measures thus may create the appearance of a significant difference between the constant terms of regressions in the two classes of states where none exist.

A simple remedy is separate estimation of supply curves in the two sets of states and a subsequent test of the equality of the subset of regression coefficients associated with comparable deterrence variables across these sets. Introduction of a dummy variable distinguishing executing and nonexecuting states in the pooled sample of states (i.e., allowance for separate constant terms) as a part of this procedure could account for either "missing variables" in the supply equation or the misspecification of execution risk measures in nonexecuting states and also provide a test for the statistical significance of the *qualitative* effects of these factors. These simple procedures can mitigate but not eliminate the potential simultaneity biases pertaining to estimation of the relationship between the murder rate and the observed risk of execution since they do not address similar difficulties within executing subsets.[37] However, the potential biases that still remain are likely to work against, not in favor of, the theorized deterrent effect of execution.[38]

A second basic problem in the estimation procedure arises from heteroscedasticity in the distribution of error terms. Since the dependent variables are estimates of mean number of offenses per population, their variances are expected to be negatively related to the relevant population base: samples from urban places. Previous cross-sectional studies have shown that the magnitude of squared residuals was negatively related to the relevant population size. In this study the heteroscedasticity postulate has been subjected to a parametric test suggested by Goldfeld and Quandt (1965). The test has been applied by dividing the sample underlying table 8 into high- and low-population states after omitting the middle k states, with k denoting the number of parameters estimated. The results prove decisively contradictory to the hypothesis of homoscedastic errors.[39] In general, the scatter of residuals derived from unweighted regressions strongly indicated the existence of a negative association between residual variances and either state, urban, or (where available) the sample populations: N_S, N_{URB}, and N_{REP}, respectively. The square roots of these population measures were used alternately as weights in deriving GLS estimates of supply functions.

2. *Empirical Counterparts of Theoretical Variables*

The many difficulties associated with use of reported crime rates as measures of the true crime rates in the population have been elaborated upon in previous studies. A problem unique to this investigation is that the reported rates in 1940 and 1950 (q^o) are based on a sample of urban areas, whereas measures of explanatory variables are based on state data. Because consistent measures of all variables could not be obtained, the percentage of the state population residing in urban areas (URB) was introduced as a "correcting factor" in the regression analysis. An independent reason for the introduction of this variable is given in Section IB.

The basic theoretical deterrence variables of interest relate to the mean perceived probabilities of conviction and execution in the population and to the anticipated cost of imprisonment. The respective empirical counterparts of these variables have been estimated via measures of objective risks of conviction and

---

[37] For example, Schuessler's ranking of 41 executing states according to their average homicide rates and execution risks over the period 1937–49 showed that states ranked in the highest quartile by their homicide rates had the second highest measure of unconditional risk of execution (see Schuessler [1952, p. 60]).

[38] The argument can be illustrated unambiguously for the following simplified system. Let

$$\ln q = a + b \ln Pe + u, \tag{I}$$

$$\ln Pe = \alpha + \beta \ln q + \gamma \ln Z + v, \tag{II}$$

where Z denotes a vector of exogenous variables. The theory predicts that $b < 0$ and $\beta > 0$. If the disturbance terms u and v are uncorrelated

$$\operatorname{plim} \hat{b} = b + \frac{\beta \operatorname{Var}(u)(1 - b\beta)}{\beta^2 \operatorname{Var}(u) + \operatorname{Var}(v)} > b,$$

where $\hat{b}$ denotes the linear least-squares estimator of b. The analysis is more ambiguous, however, when other endogenous variables are introduced in equation (I).

[39] For example, the F-ratios calculated on the basis of OLS residuals derived from equations 2 and 4 in table 8 and ordered against the urban populations in each state are $F_{39,39} = 5.506$ and $F_{29,29} = 5.125$, respectively, significant even at the 1 percent level.

execution and by the actual length of time served in prison by those released from prison. The construction of each of these measures is discussed in turn.

A desired measure of $Pe|c$ is the ratio of executions to convictions for capital murder in the relevant years. However, inspection of the data reveals that the number of executions per state in a given year was generally small even in 1940 and 1950 and exhibited large percentage changes from year to year. Since these fluctuations amount to severe "sampling errors,"[40] an average ratio of executions to convictions over a number of years would provide a more stable measure. Independent data on convictions are available only in the sample year and only for the category of all murders. Thus, measures have been computed by averaging the number of executions over a number of years preceding and including the sample year in calculating the numerator of $P^oe|c = (\bar{E}^o/C^o)$ (for alternative measures see table 2). Also, alternative measures have been constructed on the basis of executions for all crimes as well as for murder only. The former measure may be relevant because of the possibility that the infrequent executions for other crimes were imposed in connection with crimes where the likelihood of fatal injuries was high. Note that the observed effect of $P^oe|c$ includes, in principle, both deterrence and incapacitating effects. Assessment of the potential incapacitating impact of the death penalty relative to actual imprisonment (see n. 34) shows it to be considerably below even the smallest empirical estimate of the overall impact of $P^oe|c$ that has been derived in the present investigation.

The ratio of all prisoners received in state prisons who were convicted of a specific crime to the number of corresponding crimes known to have occurred in the same state can serve as an index of the unconditional risk of conviction $P^oc = (C^o/Q^o)$.[41] Since in 1940 and 1950 there is only information pertaining to urban crime rates, the absolute number of crimes Q^o is calculated as the product q^oN_S on the assumption that the reported urban rate is an unbiased estimate of the state rate in states with similar rates of urbanization.[42]

---

[40] It is for this reason and for a generally small variability in the extent of execution across states that data from the year 1960 are not included in this investigation. In that year and in a few earlier and subsequent years, the overall number of executions was about half of the corresponding datum in 1950 and about one-third of that in 1940. Moreover, in 12 out of 20 executing states in 1960 for which complete data were available, only a single execution occurred in that year with the trend in all states clearly leading toward a practical abolition of the death penalty. At such small levels of executions in each state and with an attenuated distribution of execution risk measures, the effects of errors in the estimates of execution risks are likely to dominate the comparisons across states.

[41] Because the numerator of P^oc is an estimate of the total number of persons committed to state prison and the denominator is an estimate of the number of crimes known to have occurred in a given year, their ratio can exceed unity as is the case for two states in 1950 where P^oc for robbery and assault exceeds one. A similar observation applies in one state in 1950 in connection with P^hc. For a related reason an estimate of $P^oe|c$ ($PX5$) is found to exceed unity in one state in 1940. The implicit assumption, of course, is that these instrumental measures are proportionally related to the true probabilities.

[42] As has been pointed out elsewhere, estimates of Pc derived by deflating measures of conviction by measures of reported crimes may bias the effect of P^oc on q^o in a negative direction due to the correlation between common error terms, although a spurious correlation in a positive direction is also possible due to the effect of the true levels of Pc on the incentive to report crimes (see n. 43 below). Note, however, that because the numerator of the dependent variable q^o is different from the denominator of P^oc in regressions using FBI crime data, the potential negative spurious correlation is mitigated in this analysis. For further elaboration on issues relating to measurement of P^oc and its estimated effects see Ehrlich (1974, appendix 1 in his section III).

Punishment severity is measured by the median time served in state prisons by prisoners first released from prison, T. Since the relevant theoretical variable is the *cost* of imprisonment m rather than its length, the partial elasticities of offenses with respect to T are expected to be lower than those with respect to m, especially when imprisonment terms are long (see Ehrlich 1974, p. 125). In addition, although an increase in m is expected to reduce the incentive to "enter" a criminal activity by the analysis of Section IA, longer imprisonment terms may also lead to greater rates of recidivism on the part of those released from prisons due to possible adverse effects of imprisonment on relative legitimate employment opportunities. In contrast, imprisonment entails an incapacitating effect on those imprisoned, which operates to reduce the rate of crime independently of the effect of punishment on incentives. In general, the observed effect of T, and consequently also that of P^oc, as measured in the regression analysis, is a mix of all of these effects. (For some calculations of the potential incapacitating effects of imprisonment see Ehrlich [1974, p. 103].)[43] A list of all variables used in the regression analysis is given in table 2.[44]

References

Aday, L. A., and Anderson, R. *Access to Medical Care.* Ann Arbor, Mich.: Health Admin. Press, 1975.

Becker, Gary S. "Crime and Punishment: An Economic Approach." *J.P.E.* 78, no. 2 (March/April 1968): 169-217.

Box, G. E. P., and Cox, D. R. "An Analysis of Transformations." *J. Royal Statis. Soc.* 26, Ser. B (1964): 211-43.

Ehrlich, Isaac. "Participation in Illegitimate Activities: An Economic Analysis." Ph.D. dissertation, Columbia Univ., 1970.

———. "The Deterrent Effect of Capital Punishment: A Question of Life and Death." Working Paper no. 18, Nat. Bur. Econ. Res., 1973.

———. "Participation in Illegitimate Activities: An Economic Analysis." In *Essays in the Economics of Crime and Punishment*, edited by G. S. Becker and W. M. Landes. New York: Columbia Univ. Press, 1974.

———. "The Deterrent Effect of Capital Punishment: A Question of Life and Death." *A.E.R.* 65 (June 1975): 397-417. (*a*)

---

[43] Data sources not previously cited include U.S. Bureau of the Census, *Prisoners in State and Federal Prisons and Reformatories* (1936-41, 1946) and U.S. Department of Justice (1954) for E^o; U.S. Bureau of the Census, *Prisoners in State and Federal Prisons and Reformatories* (1940) and U.S. Department of Justice (1954) for C_I^o; U.S. Bureau of the Census, *Statistical Abstract of the United States* (1945) and *U.S. Census of the Population: 1950* (1952) for the variables W and X; and U.S. Bureau of the Census, *Statistical Abstract of the United States* (1946, 1953) for N_S, NW, AGE, A_{25-34}, N_{URB}, and URB.

[44] The implicit assumption in the regression analysis is that the proportion of underreported offenses is constant across states. However, the incentive to report a crime is affected by the relevant costs and gains associated with reporting. To the extent that explanatory variables introduced in the regression also affect the incentive to report, their estimated effects would be a mix of their effects on the incentive to commit as well as to report offenses. For example, individuals may be more willing to report an offense if the probability that the offender will be apprehended and punished and the severity of punishment are high. Persons with low opportunity cost of time may be more likely to bear the costs of reporting a crime of given severity, including the costs of filing a complaint and appearing as witnesses in court. The estimated effects of P^oc, T, and X, for example, might therefore be biased in a positive direction and that of W in a negative direction relative to their true effects.

———. "Deterrence: Evidence and Inference." *Yale Law J.* 85 (December 1975): 209–27. (*b*)

Goldfeld, S. M., and Quandt, R. E. "Some Tests for Homoscedasticity." *J. American Statis. Assoc.* 60 (June 1965): 539–47.

Klebba, A. J. "Homicide Trends in the United States, 1900–74." *Public Health Reports* 90, no. 3 (May–June 1975): 195–204.

Pashigian, B. Peter. "The Hill-Burton Program: The Effects of the Federal Subsidy in Kind on the Hospital Industry." Report no. 7346, Center Math. Studies Bus. and Econ., Univ. Chicago, 1973.

Schuessler, Karl F. "The Deterrent Influence of the Death Penalty." *Ann. American Acad. Polit. and Soc. Sci.* 284 (November 1952): 54–62.

U.S. Bureau of the Census. *Prisoners in State and Federal Prisons and Reformatories.* Washington: Government Printing Office, 1936–41, 1946.

———. *Vital Statistics of the United States, 1940.* Pt. 1. Washington: Government Printing Office, 1943.

———. *Statistical Abstract of the United States.* Washington: Government Printing Office, 1945, 1946, 1953.

———. *U.S. Census of the Population: 1950.* Vol. 2. Washington: Government Printing Office, 1952.

———. *Historical Statistics of the United States from Colonial Times to 1957.* Washington: Government Printing Office, 1960.

U.S. Department of Health, Education, and Welfare, Public Health Service. *Vital Statistics of the United States, 1950.* Vol. 3. Washington: Government Printing Office, 1953.

———. *Vital Statistics of the United States, 1971.* Vol. 2. Washington: Government Printing Office, 1975.

U.S. Department of Justice, Bureau of Prisons. *National Prisoner Statistics: Prisoners Released from State and Federal Institutions.* Washington: Government Printing Office, 1951, 1952–53.

———. *National Prisoner Statistics: Prisoners in State and Federal Institutions, 1950.* Leavenworth, Kans., 1954.

U.S. Department of Justice, Federal Bureau of Investigation. *Uniform Crime Reports.* Washington: Government Printing Office, 1940, 1950–51, 1960, 1970.

Wolfgang, Marvin. "A Sociological Analysis of Criminal Homicides." *Federal Probation* 25 (March 1961): 48–55.

The Demand for Supplementary Health Insurance, or Do Deductibles Matter?

Emmett B. Keeler, Daniel T. Morrow, and Joseph P. Newhouse

RAND Corporation

A model in which the individual with median anticipated health expenditures can choose the level of deductible supplementation for his group is used to estimate the demand for supplementary insurance. Differences between individually purchased and group insurance are discussed. Claims and expenditure data are used to show that demand for supplementation of outpatient services will be small unless the tax subsidy of insurance is continued. Experience from Medicare corroborates these predictions.

A number of economists have advocated national health insurance (NHI) plans that contain a large income-related deductible (Feldstein 1971; Pauly and Kihlstrom 1971). Other proposed NHI plans have provided for moderate income-related deductibles (the 1974 administration and Kennedy-Mills Bills) or larger nonincome-related deductibles (the Long-Ribicoff Bill). But it is argued by many that such deductibles would fail to achieve their aims because consumers would simply purchase private insurance that supplemented the NHI plan and would therefore be fully covered. The National Center for Health Services Research appears to feel there is a consensus on this point; their recently issued *Program for Health Services Research* (1976) states that "it generally is believed that consumers might subvert [an income-related deductible's] intent by the

---

The authors wish to thank Willard Manning, Bridger Mitchell, Charles Phelps, Robert Roll, and Michael Rothschild for helpful comments on an earlier draft. The research reported herein was performed as part of the Health Insurance Study, a grant from the U.S. Department of Health, Education, and Welfare, Washington, D.C. The opinions and conclusions expressed are solely those of the authors and should not be construed as representing the opinions or policy of any agency of the U.S. government.

[*Journal of Political Economy*, 1977, vol 85, no. 4]

© 1977 by The University of Chicago. All rights reserved

purchase of supplementary, front-end insurance." Numerous others have also argued in this manner (Eilers 1971, p. 264–65; Evans 1973, p. 120; Cohen 1975, p. 295; McClure 1976).

This paper considers whether there would be a demand for supplementary insurance below a moderate NHI deductible of $200 per person per year or less. Other work suggests that this sized deductible could well alter the demand for medical care relative to a full coverage plan (Newhouse, Phelps, and Schwartz 1974), which makes the potential market for supplementary insurance an important issue. Our analysis suggests that demand for supplementary insurance that eliminates the cost sharing will be negligible if there is no favorable tax treatment of supplementary insurance. For the larger deductibles of the Long-Ribicoff or Feldstein plans demand for some supplementation may be substantial, but our calculations show that full supplementation would not be sought.

In deriving our results, we also develop a new theoretical proposition. Arrow (1963) has shown that the marginal value of individually purchased insurance to a risk-averse consumer rises as the deductible rises. We show that this result need not hold for group-purchased insurance. Because more than 80 percent of private health insurance is currently purchased through groups, this qualification is of practical importance.

I. Policy Significance of Supplementary Insurance

In this paper we analyze supplementary insurance that covers a moderate deductible instead of policies that reduce any coinsurance that may apply above the deductibles because we believe coinsurance has less effect on demand for outpatient services.[1] We expect little increase in demand for hospital services from a NHI program because of extensive existing coverage, but the increase in demand for office visits could be 75 percent or more if no deductible is included in a national plan (Newhouse et al. 1974). Because existing supply could not accommodate such a demand increase, there would very likely be extensive nonprice rationing of services if supplementary insurance were widespread. Moreover, full-coverage insurance eliminates any basis for price competition, whereas a moderate deductible preserves the possibility of price competition for nonhospital services.

Theory suggests that demand as a function of a deductible approximates a logistic curve (fig. 1) (Keeler, Newhouse, and Phelps 1977). If the NHI deductible were set at d_1 with the intent of restraining demand,[2] it is

---

[1] Nor do we consider demand for insurance that might cover fees charged in excess of what a public insurance plan will pay for. In effect, we assume such insurance is in force (if the public plan does not pay for everything) through our assumption that the coinsurance rate above the deductible is zero.

[2] Opponents of a NHI deductible argue that it would perversely limit demand for preventive or early care and thus increase total demand in the long run and lead to worse health outcomes. There is little, if any, evidence to either support or refute this claim. However, if true, it is possible to waive cost sharing for certain preventive services such as

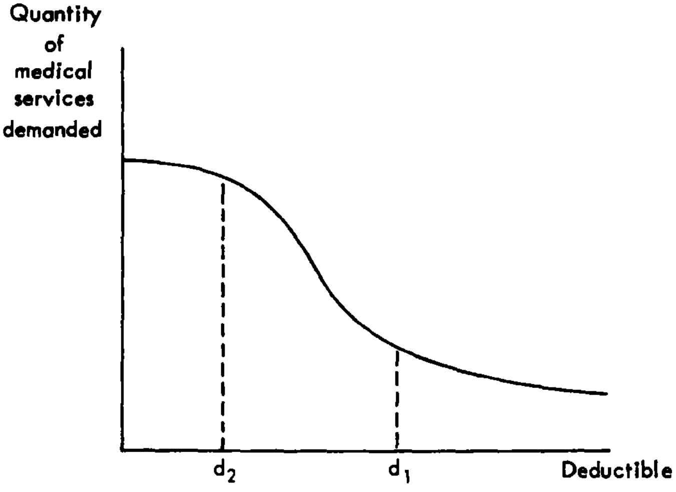


FIG. 1

important to know whether or not many persons would supplement down to d_2 or beyond. If this occurred, inclusion of a deductible in the NHI plan would not be effective in avoiding the demand increase. There is little evidence on where d_1 might lie; however, only 28 percent of the population had expenditures exceeding $200 in a national probability sample survey, which is consistent with d_1 lying near $200.

The discussion in this paper assumes that the cost-sharing provisions in a NHI plan apply to total expenditures (i.e., unreimbursed and reimbursed). The entire issue of supplementary insurance could be rendered moot by defining the national health insurance plan to apply to *unreimbursed* expenditure (just as the current medical deduction on the personal income tax does). In that case, it is extremely unlikely that there would be any demand for supplementary insurance.[3] Since none of the major

---

immunizations. The level of the deductible also determines the proportion of total expenditures for medical services which flow through the public budget for NHI Because of the opportunity cost or political cost of public expenditures, it may be desirable to shift medical expenditures from the public to private budgets. For example, increasing the annual per person NHI deductible from $100 to $200 would, we estimate, transfer $5.7 billion to private budgets.

[3] Suppose the NHI plan has a deductible of $X that applies to unreimbursed expenditures. In that case, an individual who purchases a supplementary policy that covers the first $X implies he is willing to pay the loading fee for the policy rather than incur an exposure of $X. But, of course, after purchasing this policy, he still has an exposure of $X so by assumption he would buy a policy to cover the next $X. (There is an income effect that tends to restrain the consumer from continuing in this fashion, but the income effect should be empirically negligible.) Thus, the consumer, if he supplements at all, would generally continue to supplement until he is completely covered by a private policy. However, at complete coverage the actuarial value of the private policy will amount to

NHI bills before the Congress limits public insurance to unreimbursed expenditure, nor does the Medicare program, we may proceed with our analysis.

II. A Model of Supplementation

The *loading fee* of an insurance policy is the proportion by which the premium charged to the consumer exceeds the expected insurance payments to cover the costs of selling and administering the policy and risk bearing by the insurer. The lower the loading fee, the more attractive insurance becomes to the consumer (Phelps 1973). In this section we show that even with assumptions that should favor supplementation the maximum loading fee a consumer would be willing to pay for any supplementation below a moderate deductible is generally smaller than any company could offer without a tax subsidy.

We first derive a simple formula for the largest loading fee that the consumer would pay and then use expenditure data from the 1970 survey of the Center for Health Administration Studies and the National Opinion Research Center (CHAS-NORC) to estimate the parameters in the formula (Anderson et al. 1973). To simplify our argument, we assume that (1) supplementary insurance has no effect on demand; (2) the coinsurance rate above the deductible rate is zero, and (3) insurance is purchased through a group. However, these assumptions are not critical to our results.

The effects of supplementation on demand are the reason for this analysis, so it may seem odd to assume that supplementary insurance has no effect. Nevertheless, it makes analysis much simpler, and if supplementary insurance is not bought if it has no effect on demand it should not be bought if it increases demand. This follows if there is any welfare loss to the consumer from the subsidy that supplementary medical insurance represents (see Keeler et al. 1977 for an estimate of this welfare loss).[4]

Our second assumption is that the coinsurance rate above the deductible is zero. Reduction of coinsurance should be more attractive to more consumers than reduction of the deductible because its marginal administrative costs should be relatively small (the company administering the policy covering above deductible expenses has only to increase the amount of a check it must already write) and because demand should be more

---

several hundred dollars, so that the premium will exceed the values of $\$X$ that we consider. Thus, the consumer will prefer a probability of spending $\$X$ to the certainty of spending a larger amount on a premium, and there will be no market for supplementary insurance.

[4] Because supplementation increases demand, individuals are more likely to exceed a deductible and so obtain publicly financed benefits for expenditures in excess of the deductible. However, these benefits above the deductible are available to the individual whether he purchases supplementary insurance or not and so do not affect his decision to purchase supplementary insurance.

sensitive to reductions in a moderate deductible than to reductions in coinsurance (and therefore losses due to insurance distortions are greater [see fig. 1]). For these reasons, if we show that supplementary insurance is purchased against a deductible we may assume that insurance against any coinsurance above the deductible will also be purchased. If we show the reverse, we cannot say whether any coinsurance above the deductible will be eliminated, but the critical policy point is whether supplementation will eliminate a deductible, as discussed in Section I.

The assumption of group insurance is favorable for supplementation because administrative costs (and hence loading fees) are much lower for group insurance than for individual insurance, and the problem of adverse selection (that only the most sickly want insurance) is not as severe.[5] Suppose then that individual i can decide the deductible level d for a group to which he belongs. Let the NHI plan have deductible D. If $B_i(d)$ is the benefit to i of his group having insurance with a deductible d, supplementation of the deductible is attractive if for some $d < D$, $B_i(d) > 0$. By definition,

$$B_i(d) = R_i(D) - R_i(d) + I_i(d) - P_g(d) \qquad (1)$$

where $R_i(d)$ is the risk premium or the dollar value of the risk with deductible d, $I_i(d)$ is the expected payment by the insurer, and $P_g(d)$ is the premium for the group policy. We shall explain each of these in turn.

The risk premium is the amount an individual would pay in excess of the actuarial value of his losses to avoid them. Assuming that the individual's preferences can be represented by a utility function that is separable in health and money, the risk premium can be computed from the individual's distribution of expenses $f_i(x)$ and his aversion to risk. Let $h_i = \min\,(x_i, d)$ be the random variable of out-of-pocket expenses with a deductible d, and let u be the individual's utility function for money. Pratt (1964) measured risk aversion by $r = -u''/u'$ and showed for small gambles Z that the risk premium is approximately $(r/2) \cdot \text{var}\,(Z)$. In the range of deductibles under investigation ($0–$200) and for reasonable measures of risk aversion ($r \leq .002$) this approximation is quite good.[6] Thus, we assume

$$R_i(d) = r/2(\text{var}\, h_i) = r/2 \cdot E[h_i - E(h_i)]^2. \qquad (2)$$

---

[5] If the insurance is not mandatory for all members of the group, adverse selection remains a problem (see Keeler, Morrow, and Newhouse 1976, pp. 21–22, and Rothschild and Stiglitz 1976).

[6] Two common families of utility functions used in theoretical examples are $-e^{-rx}$ and $-x^c$, $c < 0$. For both functions, the gamble Z for which they differ most from the approximation is a .5 chance of losing $200 and a .5 chance of no loss. If $r = .002$, the approximation gives the value of $.002/2 \cdot \text{var}\,(Z) = \$10$, while the assumed true value is $Eu(Z) = .5e^{-r(-200)} - .5e^{-r(0)} = -1.2459 = -e^{-r(109.93)}$, a difference of 7 cents. If $u = -x^c$, $r = (1 - c)/x$. Thus, $r = .002$ is consistent with $c = -9$, $x = 5{,}000$. In this case $Eu(Z) = u(109.94)$, a difference of 6 cents from the approximation.

We shall need the derivative $R_i'(d)$ below:

$$\begin{aligned} R_i'(d) &= r/2[E(2h_i \cdot h_i') - 2E(h_i)\cdot E(h_i')] \\ &= r\{d[1 - F_i(d)] - E(h_i)\cdot[1 - F_i(d)]\}, \end{aligned} \tag{3}$$

since $h_i'(d) = 0$ for $x < d$, $h_i'(d) = 1$ for $x > d$.[7]

The expected insurance company payment to the decisive individual, $I_i(d)$, is given by

$$I_i(d) = \int_d^D (x - d) f_i(x)\, dx + (D - d)[1 - F_i(D)]. \tag{4}$$

The premium for the policy is based on the expected payments by the insurance company to members of the group and the loading fee, θ. Thus, $P_g(d) = (1 + \theta)I_g(d)$ or

$$P_g(d) = (1 + \theta)\left\{\int_d^D (x - d) f_g(x)\, dx + (D - d)[1 - F_g(D)]\right\}. \tag{5}$$

The value of insurance to the consumer varies inversely with the effective loading fee. When the loading fee is so large that the premium is bigger than any possible payment, no one would want insurance; when the effective loading fee is minus one (as it would be if insurance payments could be used as a tax credit), insurance is free and everyone would want it. For any individual, any deductible, and any proposed supplementary reduction to $x \leq D$ we can find a break-even effective loading fee θ_x, such that the individual is indifferent between purchasing or not purchasing that reduction. For loading fees larger than θ_x supplementation will not be desirable, and for loading fees smaller than θ_x it will be desirable.

The θ_0 associated with full supplementation for the decisive individual in the group can be found by solving $B_i(0) = 0$ for θ to get

$$\theta_0 = \frac{R_i(D) + I_i(0)}{I_g(0)} - 1, \tag{6}$$

since $R_i(0) = 0$. The θ_D associated with a slight reduction in the deductible can be found from the marginal value of supplementation at D, $B_i'(D)$. From (2), (3), (4), and (5),

$$\begin{aligned} B_i'(D) = {} & -r[1 - F_i(D)][D - E(h_i)] - [1 - F_i(D)] \\ & + (1 + \theta)[1 - F_g(D)]. \end{aligned} \tag{7}$$

And so θ_D is found by solving $B_i'(D) = 0$ for θ to obtain

$$\theta_D = \frac{[1 - F_i(D)]\{r[D - E(h_i)] + 1\}}{1 - F_g(D)} - 1. \tag{8}$$

Formula (8) shows that supplementation will occur at larger loading fees (i.e., that demand for supplementation will be greater) when purchasers are more risk averse (r is large), when the deductible is large and the average loss is small [$D - E(h_i)$ is large], and when the decisive individual's probability of exceeding the deductible is high relative to the group's ($[1 - F_i(D)]/[1 - F_g(D)]$ is large).[8]

---

[7] Define $F_i(x) = \int_0^x f_i(t)\, dt$ to be the probability of individual expenditures less than x, and $F_g(x) = \int_0^x f_g(t)\, dt$ to be the proportion of group expenditures less than x.

Let $\theta_{\max} =$

$$\max_{x \leq D} \theta_x.$$

Then if the real effective loading fee is larger than $\theta_{\max}$, supplementation to any level below D will not occur. In all cases considered, our (unpublished) results show that $\theta_{\max}$ was associated with either a slight reduction in the deductible or complete elimination of the deductible (i.e., $\theta_{\max} = \max [\theta_0, \theta_D]$). In the computations of $\theta_{\max}$ below, r will be assigned several values and $d - E(h_i), f_i$, and f_g will be estimated for different groups using the CHAS-NORC data.

We now summarize the basis of our calculations briefly. A fuller explanation of the data, methods, and results is given in Keeler et al. (1976), available from the authors. We assume the group considering supplementary insurance comes from a firm whose members are representative of the under-65 nonpoor population.[9] Only expenditures likely to be covered by NHI are counted: hospital, physician services, prescription drugs, dental services for children under 18, and vision and hearing services.

The assumed NHI (annual) deductible is varied from $50 to $200 per individual and from $150 to $450 per family (in 1975 dollars). We assume risk aversion $r = .0005$ and $.002$ (in 1975 dollars). These levels are considered high and should encourage supplementation.[10]

The effective loading fee $\theta = (1 + \lambda)(1 - t) - 1$ where λ is the loading fee charged by the company and t is the tax subsidy. Since supplementary insurance would involve numerous small claims, administrative costs are high. We assume that $\lambda = 35$ percent. Employer contributions to health insurance premiums are not currently taxable. We assume a marginal tax rate of 31 percent (Mitchell and Phelps 1976). If such subsidies are not eliminated under NHI, the loading fee is thus $(1 + .35)(1 - .31) - 1 = -.07$.

---

[8] Formula (8) also shows that if the loading fee were truly θ_D, then D would be the preferred deductible level for the decisive individual. For any other θ, the preferred level of d is the one that maximizes $B_i(d)$. In particular, for individual policies or for groups of identical individuals, $f_i(x) = f_g(x)$ for all x, and the maximum occurs at the deductible d where $d - E(h_i) = \theta/r$.

[9] All major proposed national health insurance plans that include cost sharing for the general population reduce or eliminate it for the poor and continue the Medicare program for the aged.

[10] Friedman (1974) calculates $r = .0026$ from 1968 health insurance data and remarks that the value is high. Feldstein (1973) considers $r = .0007$ to be very risk averse. Since r must be deflated by the CPI, Friedman's and Feldstein's values represent .0017 and .0005, respectively, in 1975.

TABLE 1

ACTUAL AND ESTIMATED EXPENDITURES

	DOLLAR RANGE							
EXPENDITURE	0-24	25-49	50-99	100-199	200-499	500-999	1,000-1,999	2,000-∞
Actual	.34	.10	.15	.13	.12	.07	.06	.03
Estimated* . .	.36	.11	.12	.11	.12	.07	.05	.06

* The proportion of expenditures between the bounds if expenditures are distributed so that log $(x_{1975} + 1.393) \sim \Lambda(4.08, 5.45)$

We assume that the group decides whether or not to purchase supplementary insurance by majority vote (Goldstein and Pauly 1976). Since we assume everyone is equally risk averse, the individual with median anticipated expenditures is decisive for the group. The observed CHAS-NORC expenditures are roughly lognormal (table 1). We assume that individual expenditures are also lognormal with log mean that varies from person to person and a log variance 70 percent of the population log variance (Morris 1974). Thus the decisive individual has anticipated expenditures with the same median as the group, but with a smaller variance.

III. Results

Table 2 gives the premiums for full supplementation and the expected insurance payments and benefit from the risk avoided by insurance to the median individual for various groups and deductible sizes. The column marked "unsubsidized premium" assumes NHI eliminates the tax subsidy, while the next column gives the effective premium if the subsidy remains. The unsubsidized premiums are so large relative to the deductible that full supplementation in this case is clearly not attractive. Why should the median individual pay $44.70 for sure to avoid some chance of paying $50? The actuarial value of the policy is only $34.30, and the benefit from the risk avoided by insurance is only ten cents, for an r of .0005. Since the risk is linear with r, r would have to be over 100 times as large for full supplementation to be desirable.

On the other hand, if the tax subsidy is continued, the expected payments to the median individual are larger than the effective premium so there may be supplementation in this case. Still, our assumptions seem conservative, and the overall gain from supplementation is slight. This gain might well be outweighed by negative effects of demand shifts or a fixed transaction cost for buying the policy.

Table 3 gives the maximum effective loading fees consistent with purchase of supplementary insurance. The first two columns show the

TABLE 2

PREMIUMS, PAYMENTS, AND RISK AVOIDED WITH FULL SUPPLEMENTARY INSURANCE
($)

Group and Deductible	Unsubsidized Premium	Tax-Subsidized Premium	Expected Payments to Median Individual	Value of Risk Avoided by Supplementation (r = .0005)
Nonpoor adults between the ages of 18 and 65 (individual deductible):				
50	44.70	30.80	34.30	0.10
100	75.70	52.10	57.00	0.40
150	101.00	69.60	74.50	0.90
200	122.60	84.50	88.90	1.60
Children (less than 18) in nonpoor families (individual deductibles):				
50	40.20	27.70	30.40	0.10
100	63.00	43.40	45.90	0.40
150	79.30	54.60	55.90	0.70
200	91.90	63.30	63.10	1.10
Nonpoor families with head under 65 and more than one member (family deductible):				
150	171.60	118.20	132.50	0.30
300	302.80	208.60	233.70	2.30
450	411.20	283.30	315.00	6.30

loading θ_D at which the individual would be indifferent between purchasing a marginal change in his deductible and not purchasing supplementary insurance. The third and fourth columns show the θ_0 for supplementing a deductible to zero. The only entry in the table larger than the .35 loading we believe will prevail is for a $450 family deductible and a risk-aversion parameter of .002. Even in this case, θ_D falls to .35 at a deductible of $430 (not shown), so little supplementation will take place. Thus, in general, supplementary insurance will not be purchased unless there is a tax subsidy, so that the effective loading is −.07. Even in this case, insurance may not be purchased, for the reasons argued above.

Table 3 reveals an interesting fact about group insurance. Arrow (1963) showed that the marginal value of individually purchased insurance to a risk-averse consumer decreases as the deductible increases if, as we assume here, the price elasticity is zero. Table 3 illustrates that this result does not necessarily apply to group insurance in which the expenses of the decisive individual have a different distribution than the distribution of expenses for the group. For individual adults and for children the maximum loading fee *falls* as the deductible rises, the exact opposite of what would be observed in the individual case. It follows that there are loading

TABLE 3

MAXIMUM LOADING FEES, θ, COMPATIBLE WITH MANDATORY SUPPLEMENTARY DEDUCTIBLE INSURANCE

	SLIGHT SUPPLEMENTATION (θ_D)		SUPPLEMENTATION TO 0 (θ_0)	
GROUP AND DEDUCTIBLE ($)	r = .0005	r = .002	r = .0005	r = .002
Nonpoor adults between the ages of 18 and 65 (individual deductibles):				
50	.02	.05	.04	.05
100	−.02	.06	.02	.05
150	−.04	.08	.01	.05
200	−.06	.11	.00	.05
Children less than 18 in nonpoor families (individual deductibles):				
50	−.03	01	.03	.03
100	−.11	−.02	−.01	.01
150	−.16	−.03	−.04	.00
200	−.20	−.02	−.06	−.01
Noonpoor families with head under 65 and more than one member (family deductible):				
150	.07	.11	.04	.05
300	08	22	.05	.08
450	.09	.36	.05	.12

fees such that supplementation to zero would occur while marginal supplementation would not ($\theta_0 > \theta_D$).

This result occurs in the case of group insurance because the median individual's expenses are considerably less than the group average (most of the expenses are caused by a few sickly people). Thus, a negative loading fee is frequently necessary to make insurance attractive to the median individual (another result that differs from the individual case). Because of the skewed distribution of expenditures, this effect is more noticeable for larger deductibles.

Supplementary insurance that covers only hospital care would have much lower loading fees and could thus be more attractive than the broader insurance we analyzed.[11] For example, virtually all overnight stays in a hospital result in total expenditures in excess of $200; if it costs $3 to process such a claim for the hospitalization, the administrative marginal cost is less than 1.5 percent (and negligibly small if the above deductible expenditures would cause a claim to be filed anyway). These

[11] Phelps (1973) found that variation in loading fees faced by various consumers was the single most powerful predictor of amount of insurance held. We have not estimated demand for such insurance because we have no estimate of the variance across individuals in expected hospital expenditure and because we expect elimination of a moderate deductible for hospital services would have little effect on demand.

markedly lower administrative costs may explain why existing insurance for hospital services is nearly complete (over 90 percent of hospital expenditures are paid by a third party), whereas coverage for other services is much less complete (Newhouse et al. 1974).

The purchase of supplementary insurance by Medicare enrollees is sometimes cited as evidence that supplementation will occur under a NHI plan that includes cost sharing (for example, McClure 1976). However, while slightly over half of the Medicare population has purchased supplemental insurance that covers only inpatient expenditures,[12] we estimate that only 11 percent of enrollees have insurance that covers the first dollar of *physician* expenditures.[13] This 11 percent reflects the partial deductibility of insurance premiums in income taxes and subsidized coverage of retirees by some existing group insurance plans.[14] Thus, the small fraction of the aged that has eliminated the Medicare deductible for physician services and the relatively large fraction that has purchased coverage against the large losses possible with long hospital stays are consistent with our results.

IV. Conclusion

Our calculations suggest that if there is no special tax benefit for supplementary health insurance premiums demand for insurance that provides first-dollar coverage for nonhospital services is likely to be negligible. The premium for such insurance would simply be too high to make its purchase attractive.

If the present tax treatment of health insurance premiums is continued (employer-paid premiums are not taxable income and individually-paid premiums are one-half deductible for those who itemize), supplementary insurance becomes considerably more attractive. Whether it will prove attractive enough to be demanded in these circumstances is uncertain.[15] We speculate that purchases would be small even with the tax subsidy.

But if a national health insurance plan includes cost sharing, continuation of the tax treatment is difficult to defend. One rationale for cost sharing is to shift moneys for financing medical care out of the government budget; however, the tax subsidy is a form of "tax expenditure" that simply adds to the government budget, thereby defeating the aim of shifting moneys from the budget. Another rationale for moderate cost sharing is to restrain demand while protecting families against heavy financial loss. If this is the rationale, one clearly does not wish to increase demand (by a potentially large factor) by encouraging the purchase of supplementary insurance.[16]

---

[12] This result comes from unpublished data from the Current Medicare Survey that the Social Security Administration has kindly provided.

[13] The data used to make the estimate of 11 percent are taken from Peel and Scharff (1973; see Keeler et al. [1976], p. 24 for details).

[14] For example, federal employees who have been members of the Federal Employees Health Benefits Plan can continue to participate in that plan (which covers physician expenditures) after they become eligible for Medicare at a premium which is subsidized by the federal government.

[15] Feldstein and Friedman (1974) used a different model to point out the demand for insurance and hence medical care is increased by the tax subsidy.

References

Andersen, Ronald; Kravits, Joanne; Anderson, Odin; and Daley, Joanne. "Expenditures for Personal Health Services." HRA-74-3105. Washington: Government Printing Office, October 1973.

Arrow, Kenneth J. "Uncertainty and the Welfare Economics of Medical Care." *A.E.R.* 53, no. 5 (December 1963): 941–73.

Cohen, Wilbur. *Testimony before the Subcommittee on Health of the Committee on Ways and Means, House of Representatives, July 24, 1975.* Washington: Government Printing Office, 1975.

Eilers, Robert. "Comment." In *National Health Insurance: Conference Proceedings,* edited by Robert Eilers and Sue Moyersman. Homewood, Ill.: Irwin, 1971.

Evans, Robert G. *Price Formation in the Market for Physician Services.* Ottawa: Information Canada, 1973.

Feldstein, Martin S. "A New Approach to National Health Insurance." *Public Interest* 23 (Spring 1971): 93–105.

———. "The Welfare Loss of Excess Health Insurance." *J.P.E.* 81, no. 2, pt. 1 (March/April 1973): 251–80.

Feldstein, Martin, and Friedman, Bernard. "Tax Subsidies, the Rational Demand for Insurance and the Health Care Crisis." Discussion Paper no. 382, Harvard Inst. Econ. Res., September 1974.

Friedman, Bernard. "Risk Aversion and the Consumer Choice of Health Insurance Option." *Rev. Econ. and Statis.* 56, no. 2 (May 1974): 209–14.

Goldstein, Gerald, and Pauly, Mark. "Group Health Insurance as a Local Public Good." In *The Role of Health Insurance in the Health Services Sector,* edited by Richard Rosett. New York: Nat. Bur. Econ. Res., 1976.

Keeler, Emmett B.; Morrow, Daniel T.; and Newhouse, Joseph P. "The Demand for Supplementary Health Insurance, or Do Deductibles Matter?" R-1958-HEW. Mimeographed. Santa Monica, Calif.: RAND, July 1976.

Keeler, Emmett B.; Newhouse, Joseph P.; and Phelps, Charles E. "Deductibles and the Demand for Medical Services: The Theory of a Consumer Facing a Variable Price Schedule under Uncertainty." *Econometrica,* 1977 (forthcoming).

McClure, Walter. "The Medical Care System under National Health Insurance: Four Models." *J. Health Politics, Policy and Law* 1 (Spring 1976): 46.

Mitchell, Bridger M., and Phelps, Charles E. "National Health Insurance: Some Costs and Effects of Mandated Employee Coverage." *J.P.E.* 84, no. 3 (June 1976): 553–71.

Morris, Carl. "Some Results Bearing on Experimental Precision." Unpublished memorandum, August 26, 1974 (available from the authors).

[16] Note also that purchase of supplementary insurance will increase demand for benefits in excess of the deductible; that is, in the public plan. Pauly (1974) has suggested a positive tax on supplementary insurance equal to the expected additional demand in the public program. Such a tax, if enacted, would obviously further reduce the likelihood of any demand for supplementary insurance.

National Center for Health Services Research. "The Program in Health Services Research." DHEW publication no. (HRA) 76-3136. Washington: Government Printing Office, 1976.

Newhouse, Joseph P.; Phelps, Charles E.; and Schwartz, William B. "Policy Options and the Impact of National Health Insurance." *New England J. Medicine* 290, no. 24 (June 13, 1974): 1345–59.

Pauly, Mark V. "Overinsurance and the Public Provision of Insurance: The Roles of Moral Hazard and Adverse Selection." *Q.J.E.* 88, no. 1 (February 1974): 44–62.

Pauly, Mark V., and Kihlstrom, Richard. "An Analysis of Government Health Insurance Plans for Poor Families." *Public Policy* 19, no. 3 (1971): 489–521.

Peel, Evelyn, and Scharff, Jack. "Impact of Cost Sharing on Use of Ambulatory Services under Medicare, 1969." *Soc. Security Bull.* 36, no. 10 (October 1973): 3–24.

Phelps, Charles E. "Demand for Health Insurance: A Theoretical and Empirical Investigation." R-1054-OEO. Mimeographed. Santa Monica, Calif.: RAND, July 1973.

Pratt, John. "Risk Aversion in the Small and in the Large." *Econometrica* 32 (January-April 1964): 122–36.

Rothschild, Michael, and Stiglitz, Joseph. "Equilibrium in Competitive Insurance Markets: An Essay on the Economics of Imperfect Information." *Q.J.E.* 90, no. 4 (November 1976): 629–50.

Probability and Utility Estimates for Racetrack Bettors

Mukhtar M. Ali

University of Kentucky

Subjective and estimated objective winning probabilities are obtained from 20,247 harness horse races. It is shown that subjectively a horse with a low winning probability is exaggerated and one with a high probability of winning is depressed. Various hypotheses characterizing the bettors' behavior to explain the observed subjective-objective probability relation are explored. Under some simplified assumptions, a utility of wealth function of a decision maker is derived, and a quantitative summary measure of his risk attitude is defined. Attitude toward risk of a representative bettor is examined. It is found that he is a risk lover and tends to take more risk as his capital dwindles.

Several laboratory experiments (Preston and Baratta 1948; Yaari 1965; Rosett 1971) suggest that in making a decision under uncertainty low probability events are overbet and high probability events are underbet. None of these experiments was conducted under a natural environment of the decision makers. In two separate studies of actual bettings in horse races, Grifith (1949, 1961) arrived at the conclusion similar to those of Preston and Baratta and Yaari. McGlothlin (1956) repeated the Grifith (1949) study with a different set of data on horse races, and among other things he confirmed Preston and Baratta's (1948) findings. None of the Grifith and McGlothlin studies considered more than 1,500 races to estimate several probabilities. Martin Weitzman (1965) analyzed over 12,000 races to obtain a relationship between the objective probability of

---

The author wishes to acknowledge the help and ecouragement received from his wife, Julia W. Ali. She spent hours of her time in collecting the data for the analysis. Comments on an earlier draft received from J. R. Marsden and Lawrence Fisher are highly appreciated. Discussions by the participants at the Econometric and Statistics Colloquium conducted by Arnold Zellner at the University of Chicago have been beneficial in writing this paper. The comments of the editor and the referee are highly appreciated.

[*Journal of Political Economy*, 1977, vol. 85, no. 4]

© 1977 by The University of Chicago. All rights reserved.

TABLE 1

DATA DESCRIPTION

Race Track and Year	Racing Dates (*N*)	Races (*N*)	Average Daily Attendance	Average Bet per Race ($)*	Average Bet per Person in a Race ($)
Saratoga					
1970	193	1,909	3,784	23,701	6.26
1971	182	1,779	3,873	25,500	6.58
1972	187	1,829	3,393	23,706	6.98
1973	189	1,834	3,486	25,350	7.27
1974	172	1,721	3,541	25,694	7.26
Roosevelt:					
1970–71†	192	1,698	20,426	228,884	11.21
1972	154	1,355	17,014	213,082	12.52
1973	153	1,347	17,148	214,110	12.49
1974	159	1,406	15,789	216,435	13.71
Yonkers:					
1971	155	1,381	18,025	224,289	12.44
1972	145	1,268	17,258	234,837	13.61
1973	160	1,407	15,871	225,639	14.22
1974	148	1,313	15,988	227,853	14.25

NOTE.—All tracks are in New York.
* Includes all possible betting opportunities.
† Oct.–Dec. 1970, and March–Oct. 1971. In the analysis the time period is treated as the year 1971.

a horse to win and the return for a dollar bet if the horse wins. The subjective probability defined in the Grifith or McGlothlin studies is proportional to the reciprocal of this return. The estimated relationship confirms the Grifith and McGlothlin findings.

The studies of Grifith, McGlothlin, and Weitzman were limited to the thoroughbred races. The probability estimates of these studies have serious estimation errors. The estimates obtained by Weitzman are the most reliable in which a coefficient of variation of 16 percent or higher is common.[1] This can be considered large to detect even a difference of say, 20 percent, between objective and subjective probabilities. Moreover, serious technical difficulties, though ignored, arise in their grouping of observations (see Sec. I) to estimate the objective probabilities.

The first theoretical attempt to explain the observed discrepancies between objective and subjective probabilities is by Rosett (1965). Weitzman (1965) constructed a representative utility of wealth function as an implication from the relationship between objective and subjective probabilities.

In this paper, public betting behavior in 20,247 harness horse races is analyzed. The data are described in table 1. Subjective and estimated objective winning probabilities are obtained in Section I. It is found that subjectively a horse with a low probability of winning is overstated and that a high probability of winning is understated. Section II explores various hypotheses characterizing the bettors' behavior to explain the observed subjective-objective probability relation. Specifically, a theorem is proved showing a possible alternative explanation of these biases. Section III is devoted to studying the betting public's risk attitudes, especially as they are affected by the amount of money available to bet. A numerical measure of risk attitude is defined. It is found that the betting public behaves as risk lovers. The risk attitude seems to change with the amount of money (capital) available to bet. The representative bettor becomes more risk loving as his capital declines. Concluding remarks are in Section IV.

---

[1] Improved estimates can be obtained from his estimated objective probability-return relationship, but the significance of this improvement cannot be ascertained from the available data.

I. Probability Estimates

Besides the popular betting of the "place," "show," "daily double," etc., there is one regular betting opportunity in a race known as "win" bet. Betting on a horse to win pays only when the horse finishes first in the race.[2] Payoffs are according to the odds. Odds are profits per dollar bet to a successful bettor. Odds are determined from the bets made by the public, the track take, and breakage.[3] Odds for different types of bets—win, place, show, etc.—are computed separately. Suppose there are H horses numbered[4] 1, 2, . . . , h, . . . , H, and let X_h be the amount of money that has been bet for horse h to win. Then,

$$W = \sum_{h=1}^{H} X_h$$

is the total win bet in the race and is called the "win pool." Let the track take and breakage be α. Then the odds on horse h are

$$a_h = [(1 - \alpha)W - X_h]/X_h = (1 - \alpha)W/X_h - 1, h = 1, 2, \ldots, H. \tag{1}$$

---

[2] Races with "dead heat" at any finishing position are not considered.

[3] A fixed proportion of the amount bet in a race is taken out by the track before it distributes the rest to the successful bettors. This proportion is known as the track take. The breakage arises because of the following two restrictions: (a) odds cannot be below a certain minimum, and (b) odds have to be rounded downward except when the former restriction is in effect, in which case, it is rounded upward. For the races that are analyzed, all odds are rounded to 10 cents and the minimum odds are also 10 cents.

[4] Sometimes several horses are grouped, known as "entry" or "field," for a single betting interest, and the group is assigned a single number. A bet on this number is successful if one of the horses in the group is successful. Thus, if it is a win bet, the bet is successful if one of the horses in the group finishes first. Without loss of generality, this group will be taken as a single horse.

The estimates of a_h's which are reported to the public once in every minute till the race starts fluctuate over the entire betting period. However, as fluctuation is minimum in the last few minutes of the betting period when the major amount of bets are placed, and as our data are limited to only final odds, we will proceed as if the final a_h's prevailed throughout the period (for an alternative justification of this assumption see Rosett [1965]). Our data are also limited to the odds on win bets, and therefore we will confine ourselves to the win bet only and refer to it as a bet.

Following Grifith and McGlothlin, the subjective probability of horse h to win is defined as the proportion of the win pool that is bet on it,[5] (i.e., X_h/W). This can be shown to be equal to $(1 - \alpha)/(1 + a_h)$, and from (1) it can be verified that

$$\sum_{h=1}^{H} \left(\frac{1 - \alpha}{1 + a_h}\right) = 1 \tag{2}$$

where H is the number of horses in the race.

The objective winning probability of a horse is defined to be the proportion of times the horse wins when the race is repeated an infinitely large number of times. Thus, a race can be taken as a binomial trial with this objective probability for the winning outcome. In a race the total of the objective probabilities is one.

Both subjective and objective probabilities are different for horses in a race, and also they differ in different races. Let the subjective probability for a horse to win the ith race be ρ_i, the objective probability be π_i, and the odds be a_i. Then $\rho_i = (1 - \alpha)/(1 + a_i)$, where α is the track take and breakage. From our data on the odds, α can be computed using (2), and hence ρ_i can be obtained. However, we have only one observation to estimate π_i. Thus, no reasonably reliable estimate of π_i can be obtained. This estimation difficulty is inherent, though ignored in the studies of Grifith, McGlothlin, and Weitzman.

In our study the horses are grouped, and their average subjective probability is compared with an unbiased estimate of the corresponding average objective probability. Horses in the same group are identified with a single horse. Let there be N races competed in by a horse with ρ_i and π_i as its subjective and objective probabilities to win the ith race. Then its average subjective and objective probabilities are defined as

---

[5] Considering each dollar bet as a vote to win, the subjective probability is the total public support for the horse to win the race. It should be noted that this definition of subjective probability is not to imply expected net gain from betting a dollar on a horse is $-\alpha$, though the subjective expected net gain is $-\alpha$ and is the same for every horse. This subjective probability is also the reciprocal of the return to a dollar bet if the horse wins when there is no track take or breakage. Thus, when there is no possibility of confusion, these two terms, subjective probability and return, will be used interchangeably.

$\bar{p} = \sum p_i/N$, and $\bar{\pi} = \sum \pi_i/N$, respectively. When there is no possibility of confusion, these averages will be referred to as the respective winning probabilities of the horse. Once p_i's are computed $\bar{p}$ can be obtained. An unbiased estimate of $\bar{\pi}$ is $\hat{\pi} = \sum Y_i/N$, where $Y_i = 1$ if the horse wins the ith race and equals zero otherwise. As the races are considered independent binomial trials, it can be shown that $E(\hat{\pi}) = \bar{\pi}$, and var $(\hat{\pi}) = \bar{\pi}(1 - \bar{\pi})/N - \sigma_\pi^2/N$, where $\sigma_\pi^2 = \sum (\pi_i - \bar{\pi})^2/N$. In the absence of a reasonable estimate of σ_π^2, we neglect it in estimating var $(\hat{\pi})$ which is estimated as $\hat{\pi}(1 - \hat{\pi})/N$. Thus, the reported standard error of $\hat{\pi}$ which is the square root of this variance-estimate is exaggerated.

In grouping the horses we use the following criteria: (*a*) N should be as large as possible; and (*b*) σ_π^2 should be as small as possible. By choosing a large N, var $(\hat{\pi})$ is reduced so that $\hat{\pi}$ is a meaningful estimate of $\bar{\pi}$. Criterion (*b*) is necessary for a valid test of the hypothesis of subjective over- or underestimation of objective probabilities by comparing $\bar{\pi}$ with $\bar{p}$. Previous studies have grouped the horses according to their odds falling in various selected ranges. For this grouping, it is very likely to have more than one horse in a group competing in the same race. Thus, their identity with a single horse is destroyed. To avoid this possibility, such competing horses can be treated as an "entry" or "field" (see n. 4). However, the odds on this field which are to be estimated should be lower than the individual odds on these horses, and thus the field may not belong to the group from which the individual horses are drawn. Hence, not only need we estimate the odds on the field but also the groups need to be rearranged. Thus, we abandon this method of grouping.

It can be argued that the variation in the winning probabilities of the horses most likely to win the races may not be substantial, that is, σ_π^2 would be small for this group of horses. Also, for this group N would be as large as the total number of races. Hence this grouping closely satisfies criteria *a* and *b*. Similar arguments can be put forward for the horses which are second most likely to win the races. Thus, without any further information, the horses are grouped according to "favorites." The horse with the lowest odds in a race is called the first favorite; the horse with the next lowest odds is known as the second favorite, and so on. The hth favorite in a race will be called horse h,[6] and its subjective and objective winning probabilities will be denoted by $\bar{p}_h$ and $\bar{\pi}_h$, respectively.

In the following discussion, no reference is made to horses 9 and 10 because the number of races in which they competed is small. The estimates $\hat{\pi}_h$'s and their standard errors are in columns 3 and 4 of table 2, respectively. The standard errors expressed as percentages of the estimates

---

[6] A race is limited to the maximum of 10 horses, and often eight horses compete in a race. In our data, eight horses competed in 15,749 out of 20,247 races, whereas nine horses competed in only 299 races and 10 horses competed in only 71 races.

TABLE 2

PROBABILITY ESTIMATES

Horse (h) (1)	Races Competed (N) (2)	$\hat{\pi}_h$ (3)	SE ($\hat{\pi}_h$) (4)	$100\left(\frac{\text{SE}}{\hat{\pi}_h}\right)$ (5)	$\bar{\rho}_h$ (6)	$\frac{(\bar{\rho}_h - \hat{\pi}_h)}{\text{SE}(\hat{\pi}_h)}$ (7)
1	20,247	.3583	.0034	0.95	.3237	−10.29
2 . .	20,247	.2049	.0028	1.37	.2077	0.99
3 .	20,247	.1526	.0025	1.64	.1513	−0.52
4	20,247	.1047	.0022	2.10	.1121	3.45
5	20,231	.0762	.0019	2.49	.0827	3.49
6 . . .	20,088	.0552	.0016	2.90	.0601	3.01
7 . .	19,281	.0341	.0013	3.81	.0417	5.80
8	15,749	.0206	.0011	5.34	.0276	6.20
9	299	*.0033*	*.0033*	. . .	. . .	. . .
10 .	71	.0141	.0140	. . .	. . .	. . .

NOTE. Subjective probabilities for horses 9 and 10 are not shown because they are based on a very small number of observations. Corresponding objective probability estimates with their standard errors are shown to give an idea of the unreliability of these estimates.

(col. 5) vary from 0.95 percent to 5.34 percent and they can be considered small. The subjective probabilities are in column 6. Differences between subjective and estimated objective probabilities are expressed as ratios to the respective standard errors and are shown in column 7. Except for horses 2 and 3, these differences can be considered statistically significant. Every subjective probability with the exception of that of the first favorite exceeds the corresponding objective probability estimate. For the first favorite the subjective probability is significantly less than the corresponding objective probability. Thus, subjectively, a horse with a high winning probability is understated and one with a low winning probability is overstated. Qualitatively, this finding is in accord with that observed by Preston and Baratta (1948), Grifith (1949, 1961), McGlothlin (1956), and Weitzman (1965).

The present data extend over a number of years (1970–74), various tracks, and numerous race conditions. To examine the influence of some of these factors,[7] all the data were divided into 15 data sets. The previous analysis applied to each of these data sets pointed to the same conclusion as has already been reached.

II. Theoretical Explanation of the Differences between Subjective and Objective Probabilities

Despite the sampling errors and the grouping problem, the Grifith, McGlothlin, and Weitzman findings about the relationship between subjective and objective probabilities are qualitatively in accord with

[7] Some characteristics such as average daily attendance, average bet per race, average bet per race per person, the years of racing, and the racetracks are in table 1.

the present study. Grifith and McGlothlin as well as Preston and Baratta seemed to imply that the observed differences between subjective and objective probabilities are psychological. If the bettors are sophisticated the observed subjective-objective probability relation can be explained by adopting the Friedman-Savage (1948) expected utility hypothesis (EUH) if the utility functions are restricted to a properly chosen class.[8]

Rosett (1965) claims that the observed relationship is consistent with the hypothesis that bettors are rational, sophisticated, and have strong preference for low-probability-high-return bets.[9] This hypothesis is one manifestation of EUH with the properly restricted class of utility functions. To see this let us define the equilibrium betting opportunity set as the one constituting only the nondominant opportunities; that is, for any two opportunities in the set no one is preferred to the other by every bettor. Henceforth, we also maintain that an opportunity A is preferred to another opportunity B by *every* rational bettor if and only if the winning probability of A is at least as large as that of B and the return of A is larger than that of B or the winning probability of A is larger than that of B and the return of A is at least as large as that of B. Then one can prove (left to the reader):

Theorem 1.—For any two betting opportunities, A and B, A is not preferred to B by every rational bettor if and only if A is not preferred to B by every bettor with an increasing utility function who follows EUH. As can be shown from the above theorem, the equilibrium betting opportunity set as characterized by the rationality hypothesis is identical with the one obtained by adopting the EUH and restricting the utility functions to a class of increasing functions. In this case, EUH with a class of increasing utility functions is equivalent to the rationality hypothesis. The hypothesis that bettors have strong preference for low-probability-high-return bets restricts further the class of utility functions, and following Rosett (1965) it reduces the equilibrium set to the set where objective probability, π, is related to return, R, by $R = \pi^{\beta}$ where $\beta < 0$.

Rosett's postulated objective probability-return relationship was consistent with only a part of his data, and it systematically overstated the returns when the winning probabilities were smaller. In what follows we show that if the assumption of sophisticated bettors is replaced by the assumption that each bettor knows something about the objective winning probabilities but no one knows them exactly, then for a wide range of assumptions about these estimation errors the data can be consistent with the hypothesis that bettors are rational and risk neutral.

---

[8] The EUH asserts that each bettor has a unique utility of wealth function and each maximizes his expected utility to choose his preferred bet. Utility functions are unique only up to positive linear transformations.

[9] Bettors are rational in the sense that no one prefers a bet with a smaller winning probability and the same or lower return, or with a lower return and the same or lower winning probability to that available to him. Bettors are sophisticated in the sense that the objective winning probabilities are known to them.

For simplicity of exposition, let us assume that there are only two horses, H_1 and H_2, in a race. Let π_h be the objective (true) probability for H_h to win the race, ρ_h be an individual's subjective estimate of π_h, and a_h be the market-established final odds on H_h. For every individual, ρ_h's are nonnegative and $\rho_1 + \rho_2 = 1$; ρ_h's vary from person to person. Let the distribution of ρ_h be $F_h(\cdot)$ so that $F_h(\rho)$ is the proportion of individual estimates of π_h not exceeding ρ. To avoid mathematical complexity, we assume that the probability of an individual estimating π_h exactly is zero. Bettors are assumed rational, and the amount of a bet for an individual is decided upon in advance and it is the same for everyone. With the above assumptions we prove:

Theorem 2.—If π_h is the median of the distribution, $F_h(\cdot)$, and the bettors are risk neutral, the subjective probability for H_h to win, $\bar{\rho}_h = (1 - \alpha)/(1 + a_h)$ where $0 \leq \alpha < 1$ is the track take and breakage, exceeds the corresponding objective probability, π_h, if and only if $\pi_h < \pi_{h'} (h \neq h')$.

Proof.[10]—As bettors are risk neutral, the utility functions are linear; and as they are rational, horse H_h will be bet on if and only if

$$[(1 + a_h)\rho_h - 1] > \max\{0, [(1 + a_{h'})\rho_{h'} - 1]\}. \tag{3}$$

Without loss of generality, we assume $\alpha = 0$. As $\bar{\rho}_1 + \bar{\rho}_2 = 1$, $1/(1 + a_1) + 1/(1 + a_2) = 1$, so that from (3), the individual with subjective probability estimate ρ_h and $\rho_{h'}$ will bet on H_h if and only if $\rho_h > 1/(1 + a_h)$. Thus, the proportion of all the bettors wagering on H_h is $1 - F_h[1/(1 + a_h)] = 1 - F_h(\bar{\rho}_h)$, and this must equal $\bar{\rho}_h$ because bettors are assumed to bet the same amount. Hence, $\bar{\rho}_h$ can be obtained as a solution to the equation,

$$F_h(\bar{\rho}_h) = 1 - \bar{\rho}_h, \qquad h = 1, 2. \tag{4}$$

As $F_h(\cdot)$ is a monotonic function, the solution is unique.

If $\pi_1 = \pi_2$, then $\pi_1 = \pi_2 = \frac{1}{2}$, and as π_h is the median of the distribution, $F_h(\pi_h) = 1 - \pi_h$, so that $\bar{\rho}_h = \pi_h = \frac{1}{2}$. If $\pi_1 > \pi_2$, then $\pi_1 > \frac{1}{2}$ and $\pi_2 < \frac{1}{2}$. Consequently, $F_1(\pi_1) > 1 - \pi_1$ and $F_2(\pi_2) < 1 - \pi_2$; and as $F_h(\cdot)$s are nondecreasing functions, it follows from (4) that $\bar{\rho}_1 < \pi_1$ and $\bar{\rho}_2 > \pi_2$. Similarly, it can be shown that if $\pi_1 < \pi_2$, then $\bar{\rho}_1 > \pi_1$ and $\bar{\rho}_2 < \pi_2$. Hence the theorem.

Following the theorem, if $\pi_1 \neq \pi_2$, the larger objective probability is understated whereas the smaller objective probability is overstated. This is in accord with our empirical findings.

---

[10] Some of the assumptions and conditions of the theorem can be relaxed. For example, the condition that π_h is the median of the distribution, $F_h(.)$, can be weakened and the theorem holds if the larger of π_1 and π_2 does not fall below the median of the distribution of its subjective estimates.

If we want to maintain the hypothesis that bettors are sophisticated, the observed objective probability-return relation can be explained, as noted earlier, by EUH with a suitable restriction on the class of utility functions. One such attempt was made by Weitzman (1965) where it is assumed that bettors are homogeneous, each having the same utility function. The utility function is chosen so that the observed probability-return relation coincides with the expected utility indifference curve. It can be shown that bettors with such a utility function are not rational.[11] The source of this counterintuitive behavioral implication is the expansion possibility of the betting opportunity set through various combinations (parlay, martingale, etc.) of the simple bets available in individual races. Thus, this undesired implication can be avoided by restricting the betting opportunities to a single race. It is shown in the next section that the implied utility function is increasing with increasing rate; that is, the bettors are risk lovers. Hence, the observed discrepancies between subjective and objective probabilities may be due to estimation errors by the bettors who are risk neutral or because the bettors are sophisticated but they are risk lovers and behave as if the betting opportunities are limited to a single race.

III. Risk Attitudes

In what follows we assume that bettors are expected utility maximizers, sophisticated, and behave as if the betting opportunities are limited to a single race. Let Mr. B, like Mr. Avmart in Weitzman (1965), be a representative of all the bettors who establish the final odds. He may not be one of the bettors. Let $u(\cdot)$ be his utility of wealth function. In general, utility is also a function of various circumstantial factors. In what follows we abstract from these nuances and assume utility is exclusively a function of wealth. Let us now construct a representative race in which horses 1, 2, . . . , 8 of table 2 are competing. Let the true objective probability of horse h to win a race be $\hat{\pi}_h$. By definition, horse h did not have unique odds in all the races in which it competed. The representative (estimated) odds are taken to be a weighted average of all its realized odds. The weights are chosen so that the actual average net gain per dollar bet for horse h to win over all the races it competed in is the amount that would have been obtained if this weighted average of the odds had prevailed for

---

[11] The expected utility indifference curve is estimated as

$$\pi = [1.011 - 0.087 \log_{10}(1 + R)]/R,$$

where R is the return and π is the winning probability of a bet. A parlay (for the definition, see Rosett 1965) constructed from 10 races and using a betting opportunity with $R = 10$ and π as obtained from the indifference curve would have a probability better than three times the probability of a bet from the indifference curve with the same return as this parlay. Thus, if this is an indifference curve, bettors cannot be rational.

TABLE 3

UTILITY FUNCTION AND AVERAGE AND STANDARD DEVIATION OF GAINS FROM A DOLLAR BET

Horse (h) (1)	Estimated Odds ($\bar{a}_h$) (2)	$x = 1 + \bar{a}_h$ (3)	Utility Function $u(x)$ (4)	Average Gain ($) (5)	SD ($) (6)
1	1.5	2.5	.0574	−0.10	1.26
2	3.0	4.0	.1004	−0.19	1.63
3	4.4	5.4	.1348	−0.17	2.00
4	6.2	7.2	.1966	−0.24	2.30
5	8.7	9.7	.2699	−0.26	2.68
6	12.3	13.3	.3726	−0.27	3.22
7	17.7	18.7	.6028	−0.36	3.64
8 . . .	26.4	27.4	1.00	−0.44	4.27

every race. Let $\bar{a}_h$ be the estimated odds for horse h. These are reported in column 2 of table 3. For any race the objective as well as the subjective probabilities must add up to 1. In our case, these totals are, respectively, 1.01 and 0.99. Thus, the conditions are reasonably met. As these $\hat{\pi}_h$'s and $\bar{a}_h$'s are constructed from a number of races, the present race can be considered a representative of them.

With the choice of $\hat{\pi}_h$'s and $\bar{a}_h$'s, Mr. B's expected utility of betting M dollars on horse h is

$$E_h(u) = (1 - \hat{\pi}_h)u(X_0 - M) + \hat{\pi}_h u[(X_0 - M) + M(1 + \bar{a}_h)] \tag{5}$$

where X_0 is the initial capital of Mr. B. As the $\bar{a}_h$'s are the final odds, Mr. B must be indifferent between betting on any two of these eight horses. Thus, Mr B being an expected utility maximizer,

$$E_1(u) = E_2(u) = \cdots = E_8(u). \tag{6}$$

Without loss of generality, we assume that $X_0 - M = 0$, and the odds are in the units of M dollars where M is the common denominator of all the bets; and as the utility function is unique only up to positive linear transformations, we assume that $u(0) = 0$ and $u(1 + \bar{a}_8) = 1$, where $\bar{a}_8$ are the highest odds. Then from (5) and (6),

$$u(1 + \bar{a}_h) = \hat{\pi}_8/\hat{\pi}_h, \qquad h = 1, 2, \ldots, 7. \tag{7}$$

Using (7), the specific values of the function, $u(\cdot)$, are computed and reported in column 4 of table 3; $u(\cdot)$ is plotted in figure 1. All the points lie below the line joining the two chosen points $[0, u(0) = 0]$ and $[1 + \bar{a}_8, u(1 + \bar{a}_8) = 1]$. Hence, the utility function of Mr. B must be

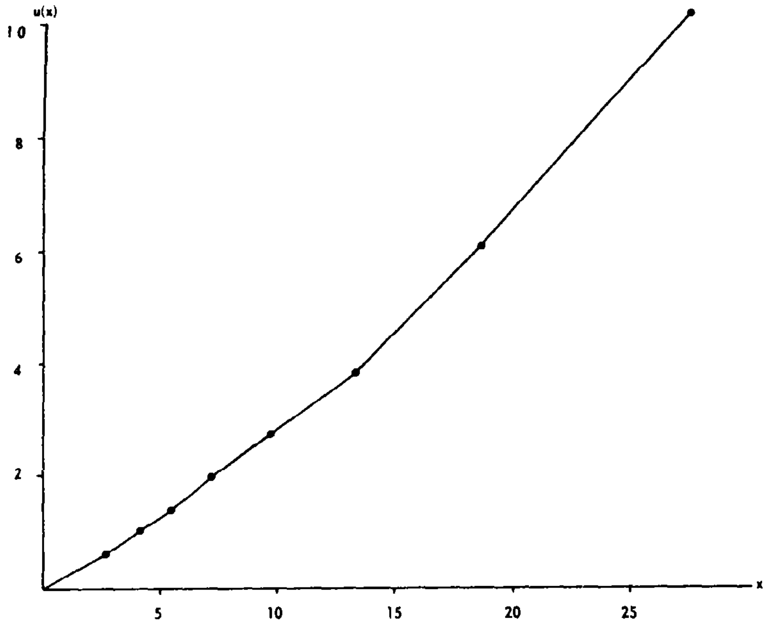


FIG. 1. Utility function

convex and he must be a risk lover.[12] These points are joined by a broken line which is the estimated utility curve.

To obtain a mathematical function representing the estimated utility curve in figure 1, $\log_{10} u(x_h)$ is regressed on $\log_{10} x_h$ where $x_h = 1 + \bar{a}_h$. The estimated relationship is given by

$$\log_{10} u(x) = \underset{(.0206)}{0.2794} + \underset{(.0210)}{1.1784} \log_{10} x, \qquad R^2 = .9981.$$

The numbers in the parentheses are the estimated standard errors. Thus, the estimated utility function is $u(x) = 1.91x^{1.1784}$. This shows $u'(x) > 0$ and $u''(x) > 0$. Though the relative risk-aversion measure of Arrow (1971), $-xu''(x)/u'(x)$, is constant, his absolute risk-aversion measure, $-u''(x)/u'(x)$, increases with wealth, implying the representative bettor takes more risk as his capital declines. This finding supports the conjecture of Markowitz (1952) for the range of the utility function under consideration.

---

[12] For each of the eight horses, the comparison of the average and standard deviation of the gains from a dollar bet computed over the 20,247 races as reported in cols. 5 and 6 of table 3 points to the same conclusion.

For a further analysis, we define a measure (new) of overall risk attitude, δ, as follows:

$$\delta = 1 - \frac{\int_a^b u(x)\, dx}{(1/b - a) \int_a^b x dx}$$

where $a = 0$, $b = 1 + \bar{a}_8$, $u(a) = 0$, $u(b) = 1$. If $\delta > 0$, the bettor is said to be a risk lover; if $\delta < 0$, he is a risk averter, and if $\delta = 0$, he is a risk neutral. The absolute magnitude of δ does not exceed 1.

For our data, $\delta = .1239$. In many respects (see table 1) the races at Roosevelt are alike to those at Yonkers. But Saratoga differs considerably from either Roosevelt or Yonkers. One striking difference is that the average bet per person per race at Saratoga is about half as much as that at either Roosevelt or Yonkers. This can be interpreted as the representative bettor at Saratoga having smaller betting capital than the one at Roosevelt or Yonkers. To investigate its effects on the risk attitude, δ values are computed from the data corresponding to each of these tracks. These are respectively, .1742, .1009, and .0652. It seems the representative at Saratoga tends to take more risk than the one at Roosevelt who in turn takes more risk than the one at Yonkers. However, the difference in attitude toward risk is less pronounced between Roosevelt and Yonkers. This can be taken as extra evidence supporting the observation that the representative bettor with less capital tends to take more risk.

To investigate further the effects of capital on risk attitude, we note that out of every dollar bet, the track take and breakage are 18¢. Thus, the total capital available to bet declines through the races during a racing night. There are usually nine races in a racing night; sometimes there are 10 or even 11, but let us proceed as if there are 9 in a racing night. If we assume that 10 percent of the money available to bet at the beginning of a race is actually put down as a bet on the race, and the track take and breakage are 18 percent (average α over all the races is .180002), then the total money available to bet at the last race will be 14 percent less than it is at the beginning of the first race. We are assuming that all the bettors arrive at the first race and stay until the last. As there are a considerable number of late arrivals, we compute δ for the first two and the last. These are, respectively, .2228, .2038, and .3040; δ values are practically the same for the first two races but are considerably larger for the last. This is in accord with our previous finding about the effect of a change in capital on the representative's risk attitude.

IV. Conclusion

The subjective winning probabilities are obtained as an average of the proportion of the total money bet on a horse. Horses 1, 2, . . . , 8 are defined as first, second, . . . , eighth favorites in any race. The objective winning probability is computed as the proportion of times a horse finishes first out of all the races in which it competes. The data consist of 20,247 races from three tracks and over a period of 5 years. It is shown that subjectively a horse with a high objective probability of winning is understated and one with a low objective probability of winning is overstated. The pattern seems to be consistent over the years of the sample period, among the chosen tracks, and under various race conditions. Theories explaining these discrepancies between objective and subjective winning probabilities are explored. It is shown that if bettors are risk neutral but not sophisticated, the market mechanism may generate the observed objective-subjective probability relationship.

Maintaining the assumption that bettors are sophisticated, have the same utility function, and behave as if the betting opportunities are limited to a single race, the representative utility of wealth function is constructed. Identifying this utility function with a representative bettor, it is found that he is a risk lover. A measure of risk attitude, δ, is defined which varies from -1 to 1. If $\delta > 0$, the person is a risk lover; if $\delta < 0$, he is a risk averter, and $\delta = 0$ means he is a risk neutral. The representative is found to have $\delta = .1239$. The effects of capital on risk attitude are investigated, and it seems that the more capital the representative has, the less he tends to be a risk lover.

References

Arrow, K. J. *Essays in the Theory of Risk-Bearing.* Amsterdam: North-Holland, 1971.

Friedman, M., and Savage, L. J. "The Utility Analysis of Choices Involving Risk." *J.P.E.* 56, no. 4 (August 1948): 279–304.

Griffith, R. M. "Odds Adjustments by American Horse-Race Bettors." *American J. Psychology* 62 (April 1949): 290–94.

———. "A Footnote on Horse Race Betting." *Transactions Kentucky Acad. Sci.* 22 (1961): 78–81.

McGlothlin, W. H. "Stability of Choices among Uncertain Alternatives." *American J. Psychology* 69 (December 1956): 604–15.

Markowitz, H. "The Utility of Wealth." *J.P.E.* 60, no. 2 (April 1952): 151–58.

Preston, M. G., and Baratta, P. "An Experimental Study of the Auction-Value of an Uncertain Outcome." *American J. Psychology* 61 (April 1948): 183–93.

Rosett, R. N. "Gambling and Rationality." *J.P.E.* 73, no. 6 (December 1965): 595–607.

———. "Weak Experimental Verification of the Expected Utility Hypothesis." *Rev. Econ. Studies* 38 (October 1971): 481–92.

Weitzman, M. "Utility Analysis and Group Behavior: An Empirical Study." *J.P.E.* 73, no. 1 (February 1965): 18–26.

Yaari, M. E. "Convexity in the Theory of Choice under Risk." *Q.J.E.* 79 (May 1965): 278–90.

The Variability of Expectations in Hyperinflations

Mohsin S. Khan

International Monetary Fund

A problem in the use of adaptive expectations as a mechanism for generating expectations of inflation is the assumption that the speed with which individuals revise their expectations is constant. The purpose of this paper is to estimate the Cagan model of the demand for money for six hyperinflation countries allowing the coefficient of expectations (which measures the response to the error between the actual and the expected rate of inflation) to vary with the level and the change in the rate of inflation. The results indicate considerable support for this particular hypothesis.

I. Introduction

The mechanism of "adaptive expectations," introduced by Cagan (1956) in his study on the monetary aspects of hyperinflations, according to which individuals form expectations about the future behavior of some economic variable, has become very popular in the literature on the behavior of the demand for money under inflationary conditions.[1]

The appeal of the adaptive-expectations model has been readily recognized, since it is a fairly simple way of representing what may be a complex set of factors behind the formation and revision of expectations. However, as Cagan was careful to point out, though this method does yield good results, it is nevertheless only an approximation. In order for the simple adaptive-expectations model to be a good representation of the true phenomenon, one requires certain assumptions which may or may not be

---

The views expressed in this paper are my sole responsibility. I am grateful to Bijan Aghevli, Phillip Cagan, Jacob Frenkel, Morris Goldstein, and Michael Mussa for extremely helpful comments. The suggestions of the referee also led to a substantial improvement in the paper.

[1] See Deaver (1970), Diz (1970), Hu (1971), Silveira (1973), and Khan (1976). This method of generating expectations has also been used extensively in other areas.

[*Journal of Political Economy*, 1977, vol. 85, no. 4]
© 1977 by The University of Chicago. All rights reserved.

satisfied. One particular assumption that Cagan himself noted explicitly as not being readily met under inflationary conditions was that expectations of inflation were formed by individuals looking only at the gap between the actual and expected rates of inflation and revising their expectations according to a fixed multiple of the forecast error, without any regard to the level of inflation.[2] It is certainly more plausible to assume that the public would become more experienced and revise its expectations of inflation faster when inflation was higher, since more intense inflation would alert it to the dangers of slow adjustment. In addition to the inflation level, the variability of inflation may be a factor in the revision of expectations. If the inflation rate is more variable, it would presumably increase the degree of uncertainty, and individuals would tend to change their expectations of inflation at a faster rate.[3] We would therefore expect that the coefficient of inflationary expectations (which measures the public's response to the error between the actual and expected rate of inflation) would be larger, the higher and more variable the rate of inflation.

There is some empirical evidence on this variability of the coefficient of expectations in hyperinflation. Cagan concluded that it was this fact that led the models fitted by him to tend to overpredict real-cash balance holdings at the beginning of the sample period and underpredict toward the end. A constant-expectations coefficient, which is in a sense an average value of the coefficient over the period, thus implies too large an effect of expected inflation on real-cash balances in the initial phase of the hyperinflation and too small an effect in the later phases. By dividing the time period and reestimating the models for the subperiods, Cagan observed a tendency for the coefficient of expectations to be larger in the later period.[4] Since the rate of inflation increased over time, it could be concluded that the coefficient varied positively with inflation. Furthermore, though less directly, the significant degree of autocorrelation in the residuals of the estimated models also provides evidence of the presence of a varying coefficient (see Cagan 1956, p. 48; Barro 1970; and Khan 1975).

The purpose of this paper is to reestimate the Cagan model for six hyperinflation countries, allowing the coefficient of inflationary expectations itself to vary with inflation, and also test for the effect of uncertainty on this coefficient.[5] This way of framing the hypothesis will allow us to test explicitly in a relatively simple way whether the coefficient varies over the hyperinflation in some systematic fashion and whether the resultant misspecification induced by ignoring this variability is responsible for the autocorrelated residuals.

---

[2] Other assumptions are required if the model is formulated in continuous time. For a discussion of these, see Foley and Sidrauski (1971) and Khan (1976). See also Frenkel (1975) for reasons why the Cagan expectations model may be inconsistent with the empirical evidence on the short-run effects of an increase in the rate of monetary expansion. A hyperinflation model using a general rational-expectations framework has been studied by Sargent and Wallace (1973).

[3] For a discussion of the role of uncertainty about inflation in the context of demand-for-money function, see Frenkel (1977).

[4] For all of the countries except Germany (the second Hungarian inflation was excluded due to an inadequate number of observations). See Cagan (1956), sec. 4. A similar test by Silveira (1973) for the case of Brazil also indicated that the coefficient of expectations increased over time.

The plan of the rest of the paper is as follows: Section II contains a description of the basic Cagan model and discusses the exact adjustment that is to be made to it. The parameter estimates are reported in Section III; and the implications of these results, together with a summary, are contained in the concluding section.

II. Specification of the Model

In the Cagan framework, the demand for real-money balances in period t is specified as a function of the expected rate of inflation in the same period. The equation in semilogarithmic form is written as

$$\log (M/P)_t = \gamma + \alpha\pi_t^* + u_t, \tag{1}$$

where M_t = stock of money balances, P_t = level of prices, and π_t^* = expected rate of inflation. Here u is a random-error term, and α is the (semi-)inflation elasticity of real-money balances (expected to be negative.)[6]

Variable π_t^* is generated by an adaptive-expectations scheme in which the expected rate of inflation is revised per period of time in proportion to the difference between the actual rate of inflation in period t and the rate of inflation that was expected in the previous period:[7]

$$\Delta\pi_t^* = \beta_t(\pi_t - \pi_{t-1}^*) \qquad 1 > \beta > 0, \tag{2}$$

where π_t is the actual rate of inflation, $\log (P_t/P_{t-1})$, and Δ is a difference operator, $\Delta\pi_t^* = \pi_t^* - \pi_{t-1}^*$. The β symbol represents the coefficient of expectations. In general, this coefficient is taken to be constant; however, I have written it with a time subscript to indicate that it can be variable.

---

[5] The Hungarian inflation of September 1945–February 1946 was excluded from the sample due to the limited number of observations available. Cagan also excluded this case from his test of the variability.

[6] For a discussion and test of whether the appropriate functional form is semilogarithmic or double logarithmic, see Frenkel (1977).

[7] The formulation differs slightly from the original Cagan specification, as it is specified in discrete rather than continuous time. In the latter form, it would be specified as

$$\frac{d}{dt}\pi^*(t) = \beta(t)[\pi(t) - \pi^*(t)] \qquad \beta > 0. \tag{2a}$$

Eq. (2) is one possible approximation to eq. (2a). It is possible that this particular approximation may yield somewhat different parameter estimates from the ones obtained by Cagan.

If, as discussed in Section I, the speed at which individuals revise their expectations depends on the rate of inflation and its variability, we can specify the following function for the coefficient of expectations:

$$\beta_t = \beta_0 + \beta_1|\pi_t| + \beta_2\tilde{\pi}, \tag{3}$$

where β is a linear function of the absolute rate of inflation ($|\pi_t|$) and the variability of inflation ($\tilde{\pi}$). Since it is hypothesized that both a higher absolute level and greater variability of inflation would lead to an increase in the coefficient of expectations (and thus shorten the lag in the adjustment of expectations), it is expected that β_1 and β_2 will be positive.[8]

The choosing of an appropriate measure of the variability of inflation, $\tilde{\pi}$, is essentially arbitrary. For this model variability is defined as simply the absolute change in the rate of inflation, $|\Delta\pi_t|$.[9] We can therefore rewrite equation (3) as

$$\beta_t = \beta_0 + \beta_1|\pi_t| + \beta_2|\Delta\pi_t|. \tag{4}$$

The reason for considering the absolute values is to ensure that β is independent of the direction of changes in prices and inflation. This makes the model equally applicable to periods of declining prices and negative changes in inflation.

Although equation (4) is proposed on the grounds of a type of a learning process and uncertainty, it is interesting to note that it can also be derived from a rational-expectations model outlined by Mussa (1976).[10]

In the case that neither $|\pi_t|$ nor $|\Delta\pi_t|$ has any influence on β, that is, β_1 and β_2 are equal to zero, the constant term in equation (4), β_0, will be equal to β. The test of the constancy hypothesis is then simply to test whether β_1 and β_2 are different from zero when the models are estimated.[11]

A clearer implication of the use of equation (4) can be seen if we substitute it into equation (2) to obtain

$$\Delta\pi_t^* = \beta_0(\pi_t - \pi_{t-1}^*) + \beta_1|\pi_t|(\pi_t - \pi_{t-1}^*) + \beta_2|\Delta\pi_t|(\pi_t - \pi_{t-1}^*), \tag{5}$$

---

[8] Note that it is implicitly assumed that uncertainty can be measured by the variation in inflation. It should be recognized, however, that there may be situations when a greater variability of inflation need not necessarily imply a greater degree of uncertainty. For example, greater variation in inflation, if fully anticipated, would involve less uncertainty than would a smaller, unanticipated variation. I am indebted to Jacob Frenkel for pointing this out.

[9] Since this definition assumes that the change in expectations is affected only by the current change in inflation, a moving-average standard deviation of inflation was also used as an alternative measure of variability. Furthermore, one- and two-period lags on $|\Delta\pi_t|$ were also experimented with. The results from these alternative definitions are discussed in Section III.

[10] If in the Mussa (1976) model the coefficient of adjustment is made a linear function of the equilibrium inflation rate, then solving one obtains an equation that relates the coefficient of adjustment to both the level and the first derivative of inflation.

[11] Functions similar to (3) have also been used by Goldstein and Khan (1976) in a study of imports and by Koot (1975) for the demand for money.

which differs from the standard adaptive-expectations model (eq. [2]) in having the two additional terms, and these two terms introduce nonlinearities into the model's specification. The first of these nonlinearities implies that the adjustment in expectations will be greater if inflation is, say, 100 percent rather than 10 percent, even though the gap between the actual and expected rates of inflation may be the same. Similarly, the second nonlinearity implies that adjustment would be greater if inflation rises from 10 percent to 20 percent than if it rises to 15 percent, again with a given gap.

Substituting (5) into (1) and eliminating the unobservable variable π_t^*, we obtain the following equation for real-money balances:

$$\begin{aligned}\log (M/P)_t &= \gamma^* + (\beta_0 + \beta_1|\pi_t| + \beta_2|\Delta\pi_t|)\alpha\pi_t \\ &\quad + (1 - \beta_0 - \beta_1|\pi_t| - \beta_2|\Delta\pi_t|) \log (M/P)_{t-1} + u_t^*,\end{aligned} \tag{6}$$

where $\gamma^* = (\beta_0 + \beta_1|\pi_t| + \beta_2|\Delta\pi_t|)\gamma$ and $u_t^* = [u_t - (1 - \beta_0 - \beta_1|\pi_t| - \beta_2|\Delta\pi_t|)u_{t-1}]$. This error term, u_t^*, now contains a moving-average process; and furthermore, due to the appearance of $|\pi_t|$ and $|\Delta\pi_t|$, which may grow with time, there is a possibility of heteroscedasticity as well.

In order to take the moving-average problem in u_t^* into account, a first-order autoregressive process for the errors was specified, as follows:

$$u_t^* = \rho u_{t-1}^* + \varepsilon_t \tag{7}$$

where ρ is the coefficient of autocorrelation, $|\rho| < 1$, and ε is assumed to have classical properties.[12]

The assumption of a constant variance in the residuals was checked by using a somewhat qualitative method. Regression equation (6) was fitted on a short segment of the first 10 observations and this segment moved along the time series. An examination of the values of the mean-squared residual from each regression will give an indication of whether the variance is changing noticeably.[13]

One point to note is that equation (6) basically relates $\log (M/P)_t$ to current and past values of prices; and with β_1 and β_2 equal to zero, it would be possible, as Jacobs has shown in a recent paper (1975), to rewrite the equation with the same dependent variable but with the current and past values of money balances as the independent variables. The two equations would be mathematically equivalent but might give different results when estimated (see Jacobs 1975). With β_1 and β_2 not equal to zero, it is extremely difficult to derive analytically a function equivalent to (6) that has only money balances (current and lagged) as explanatory variables. It is partly for this reason that the choice was made to work with an equation of the form of (6).[14]

---

[12] This procedure essentially assumes that the first-order moving-average process can be approximated by an autoregressive process of the same order. See Hendry and Trivedi (1972) for some simulation experiments on this question.

[13] Using this type of informal test was due to the nonlinear character of eq. (6) and the limited number of observations.

III. Results

Equation (6) subject to equation (7) was estimated using a combination of nonlinear least squares and the search method described by Dhrymes (1969). While estimates of the economic parameters, γ^*, α, β_0, β_1, and β_2, were obtained through the nonlinear estimation procedure, a grid search within the bounds .0 and 1.0, using steps of .1, was made for the estimate[15] of ρ. The estimate of ρ that yielded the smallest standard error in equation (6) was chosen as the appropriate estimate. Since the Durbin-Watson statistic is biased in models involving lagged dependent variables, the estimated value of ρ will provide a test of autocorrelation.

In table 1 are shown the estimates of the six parameters from the equations estimated for Austria, Germany, Greece, Hungary, Poland, and Russia.[16] Below the estimated values of the parameters are presented their respective t-values. Also shown are the coefficients of determination, R^2, the standard errors of the regression, and the Durbin-Watson test statistics.

The results show that the inflation elasticity, α, has the correct (negative) sign and is significantly different from zero at the 1 percent level in all cases. In general we observe that the estimates of the elasticity, especially in the cases for Austria, Greece, and Hungary, are smaller than those obtained by Cagan.[17]

The estimates of β_0 (which correspond to the coefficients of expectations in standard adaptive-expectations models) also have the correct (positive) signs and are significantly different from zero in all the cases. What is particularly interesting about this set of estimates is that the coefficients—ranging from 0.129 for Austria to 0.315 for Greece—are on average larger for the group than those obtained by Cagan.[18]

---

[14] Recent work by Frenkel (1977) and Khan (1977) does not indicate any particular bias involved in using a formulation such as (6). See also Cagan and Kincaid (1977).

[15] The resulting estimates of these parameters all have maximum likelihood (and therefore asymptotic normal) properties.

[16] The data used are all taken from Cagan (1956), appendix B. Variable $|\pi_t|$ is defined as $|\log P_t - \log P_{t-1}|$ for Greece, Hungary, and Russia and as $|\log P_{t-1} - \log P_{t-2}|$ for the others. Variable $|\Delta\pi_t|$ is defined as $|\pi_t - \pi_{t-1}|$ for Germany, Greece, and Poland and as $|\pi_{t-1} - \pi_{t-2}|$ for the remainder. The choice between the definitions was based on the value of the R^2 obtained from the equation. For both variables, the different definitions were used in order to introduce the possibility of lags into the behavior of the coefficient of expectations.

[17] The difference in the estimates of the elasticity could also be due to the fact that Cagan approximates the continuous-time model while we are working directly with a discrete-time version. Since the Cagan approximation involves a multiplicative factor, $1 - e^{-\beta}$, which is ignored in our estimates, some difference in the values of the elasticity can be expected.

[18] In the original Cagan estimates, the values for β range from 0.05 (Austria) to 0.35 (Russia).

TABLE 1

Demand for Money
(with Variable Coefficient of Expectations)

Country	Period	γ^*	α	β_0	β_1	β_2	ρ	$\bar{R}^2$	SE	D-W
Austria	March 1921–August 1922	−0.122	−4.496	0.129	0.125	0.074	.00	.986	0.026	2.26
		(30.29)	(49.09)	(29.87)	(9.20)	(3.39)				
Germany...........	November 1920–July 1923	−0.122	−4.343	0.178	−0.015	0.041	.50	.987	0.024	1.53
		(21.94)	(33.96)	(21.18)	(0.64)	(2.35)				
Greece	March 1943–August 1944	−0.408	−1.605	0.315	−0.476	0.587	.00	.986	0.050	1.42
		(25.25)	(8.27)	(26.24)	(16.83)	(15.78)				
Hungary	September 1922–February 1924	−0.057	−1.899	0.303	1.042	0.083	.90	.941	0.021	1.94
		(54.56)	(80.46)	(39.16)	(34.93)	(2.18)				
Poland...	June 1922–November 1923	−0.560	−2.535	0.308	0.022	0.207	.10	.981	0.030	1.93
		(36.98)	(48.53)	(36.10)	(3.63)	(20.65)				
Russia	February 1922–January 1924	−0.506	−3.608	0.245	0.023	0.103	.00	.959	0.046	1.75
		(16.94)	(8.27)	(17.59)	(0.36)	(3.48)				

Note.—Values reported in parentheses are t-ratios

The results are mixed for parameter β_1. It has a positive sign in five cases, though it is significantly different from zero only in the equations for Austria, Hungary, and Poland. For Germany and Russia one would conclude that the absolute rate of inflation had no effect on the behavior of the coefficient of expectations. Although the parameter is significant in the equation for Greece, its negative sign is not acceptable from a theoretical point of view, as it would imply that the higher the absolute rate of inflation, the slower the adjustment of expectations to the actual rate.

The effect of the absolute change in the rate of inflation on expectations, as measured by coefficient β_2, is fairly clear. The parameter has a positive sign and is significantly different from zero in all six estimated equations. The evidence therefore is consistent with the hypothesis that the speed of adjustment is influenced by the extent to which inflation is changing.

It does appear from our results that the variability of inflation has a more consistent influence on the change in expectations than does the level. The coefficient β_2 was not only significantly positive in all cases but was larger in value than β_1 in the equations for Germany, Greece, Poland, and Russia. This result is not necessarily in conflict with the Cagan conclusion that the change in expectations increases with the level of inflation, since his test merely established that over the hyperinflation the coefficient of expectations increases in size. While Cagan attributed this increase to the increase in inflation, the results here indicate that it was the increasing variability of inflation that was primarily responsible.[19]

There is some evidence of autocorrelation in the equations for Germany, Hungary, and to a lesser extent for Poland. In all other cases, the standard error of the regression equation was minimum when ρ was equal to zero. The fits of the equations, as measured by the $\bar{R}^2$s, are generally very good. An examination of the mean-squared residuals obtained from the moving regressions did not indicate any dramatic departures from the hypothesis of constant-error variance. It seems therefore that there is no serious problem with heteroscedasticity.

From the estimated parameters in table 1, the mean time lags (i.e., the time it takes for 66 percent of the effect to be completed) in the adjustment of expectations to the actual rate of inflation were calculated. We should expect that the incorporation of a variable expectations coefficient would lead to a shortening of the mean time lag. The lag is approximately

---

[19] The model was also estimated using only $|\pi_t|$ and then only $|\Delta\pi_t|$ in the equation for the coefficient of expectations. The results were very similar to those obtained using both these variables together. Estimating the model with variability of inflation defined as the standard deviation of inflation (in absolute value) calculated from a 6-month moving average yielded a positive value for β_2 for Hungary and Poland and a negative value for the rest This result does indicate that the conclusion about the role of variability depends on how the measure of variability is defined.

TABLE 2

AVERAGE TIME LAGS (MONTHS)

Country	Cagan Model*	Adjusted Model†
Austria	20.00	5.20
Germany	5.00	4.30
Greece	6.67	2.58
Hungary	10.00	1.26
Poland	3.33	2.92
Russia	2.86	4.00

* Calculated from Cagan (1956) estimates.
† Calculated from table 1 estimates.

equal to

$$L \simeq \frac{1}{\beta_0 + \beta_1|\overline{\pi}| + \beta_2|\overline{\Delta\pi}|}, \tag{8}$$

where L is the mean time or average lag and $|\overline{\pi}|$ and $|\overline{\Delta\pi}|$ are the steady-state values[20] of $|\pi|$ and $|\Delta\pi_t|$. For purposes of comparison, the mean time lags from the Cagan estimates were also calculated. If β is the coefficient of expectations,[21] the mean time lag in the Cagan model is calculated as $1/\beta$.

Both sets of average lags are shown in table 2. It is quite clear from these results that in general the average lag calculated from table 1 is shorter than the corresponding lag using the Cagan estimates.[22] Clearly, allowing for the influence of the level and changes in inflation speeds up the revision of expectations.

In summary, the results of table 1 indicate considerable support for the hypothesis that the level and variability of the rate of inflation have an effect on the change in expectations of inflation. When the residuals of the estimated models were examined, two important results emerged. First, there was considerably less autocorrelation than was present in the Cagan model; and second, there was no systematic tendency for the model to overpredict the stock of real-money balances at the beginning of the sample period and underpredict toward the end.[23] Since both autocorrelation and systematic predication errors are taken by Cagan as being possibly caused by a varying expectations coefficient, we can conclude that the present model specification is consistent with that hypothesis.

---

[20] In calculating the mean time lag, $|\overline{\pi}|$ and $|\overline{\Delta\pi}|$ were set equal to their respective average values over the sample period.

[21] The values of β were taken from Cagan, table 3, for all the countries.

[22] It is not clear why this should be the case, but it is possible that the value of β used in table 2 for the Cagan model case is an overestimate and thus the mean time lag for the constant-expectations model is underestimated. For example, if one estimates the Cagan model for Russia, one obtains a value of β equal to 0.260, which gives an average lag of 3.85 months. This value is closer to (though still smaller than) the adjusted model mean lag of 4.00 months. See also Jacobs (1975), table 1, for an estimate of β substantially smaller than the value obtained by Cagan.

[23] In order to test whether the model predictions were systematically biased, the ratio of the actual values to the fitted values of the dependent variables were plotted. In all six cases, the ratio deviated randomly about the value of unity. This of course follows from the fact that residuals were observed to be randomly distributed.

IV. Conclusion

A problem in the use of adaptive expectations as a mechanism to generate expectations of inflation is the assumption that the speed with which individuals revise expectations is constant. In conditions of hyperinflation, it has been shown by Cagan, this assumption is clearly not satisfied.

The purpose of this paper has been to reformulate the adaptive-expectations model in such a way as to incorporate explicitly the hypothesis that the gap between the actual and expected rates of inflation determines how expectations are formed, but the level and variability of the actual rate of inflation have an influence on the revision of expectations. This was done by making the coefficient of expectations a linear function of the absolute level and absolute change in the rate of inflation.[24] The higher the level and greater the changes in the rate of inflation, the faster individuals would revise their expectations.

Utilizing the adjusted expectations function, the demand for real-money balances was estimated for six of the countries originally studied by Cagan. In terms of goodness of fit, we found that there was not much to choose between our model and the Cagan model. However, the results of our model did indicate considerable support for the hypothesis that the coefficient of expectations varied in a particular way. On closer examination of the results we found two particular differences between them and Cagan's own estimates. First, the impact of expected inflation on real-money balances was generally lower than Cagan's estimates would indicate. A constant coefficient of expectations seems to result in an over-estimate of the inflation elasticity.[25] Second, the time lag between the adjustment of expectations to the actual rate of inflation was smaller than that obtained by Cagan. Both these results point to the usefulness of incorporating variable expectations into the model of the demand for money under conditions of rapidly rising prices.[26]

References

Barro, Robert J. "Inflation, the Payments Period, and the Demand for Money." *J.P.E.* 78, no. 6 (November/December 1970): 1228–63.

---

[24] Cagan has argued that the appropriate functional form may well be a step function. However, the results here indicate that the actual movement in the coefficient of expectations may be approximated by a linear function.

[25] With the reservation expressed in n. 17 above that the results here are not strictly comparable with the Cagan estimates.

[26] This type of model has also been applied with some success to other countries with relatively high rates of inflation (see Khan 1976).

Cagan, Phillip. "The Monetary Dynamics of Hyperinflation." In *Studies in the Quantity Theory of Money,* edited by Milton Friedman. Chicago: Univ. Chicago Press, 1956.

Cagan, Phillip, and Kincaid, G. "Jacob's Estimates of the Hyperinflation Model—Comment." *Econ. Inquiry* 15 (January 1977): 111-18.

Deaver, J. V. "The Chilean Inflation and the Demand for Money." In *Varieties of Monetary Experience,* edited by D. Meiselman. Chicago: Univ. Chicago Press, 1970.

Dhrymes, P. J. "Efficient Estimation of Distributed Lags with Autocorrelated Errors." *Internat. Econ. Rev.* 10 (February 1969): 47-67.

Diz, A. C. "Money and Prices in Argentina, 1935-1962." In *Varieties of Monetary Experience,* edited by D. Meiselman. Chicago: Univ. Chicago Press, 1970.

Foley, D., and Sidrauski, M. *Monetary and Fiscal Policy in a Growing Economy.* New York: Macmillan, 1971.

Frenkel, Jacob A. "Inflation and the Formation of Expectations." *J. Monetary Econ.* 1 (October 1975): 403-21.

———. "The Forward Exchange Rate, Expectations and the Demand for Money: The German Hyperinflation." *A.E.R.* (1977), in press.

Goldstein, Morris, and Khan, M. S. "Large versus Small Price Changes and the Demand for Imports." *International Monetary Fund. Staff Papers* 23 (March 1976): 200-225.

Hendry, D. F., and Trivedi, P. K. "Maximum Likelihood Estimation of Difference Equations with Moving Average Errors: A Simulation Study." *Rev. Econ. Studies* 39 (April 1972): 117-45.

Hu, T. "Hyperinflation and the Dynamics of the Demand for Money in China, 1945-1949." *J.P.E.* 79, no. 1 (January/February 1971): 186-95.

Jacobs, R. L. "A Difficulty with Monetarist Models of Hyperinflation." *Econ. Inquiry* 13 (September 1975): 337-60.

Khan, Mohsin S. "The Monetary Dynamics of Hyperinflation: A Note." *J. Monetary Econ.* 1 (July 1975): 355-62.

———. "Variable Expectations and the Demand for Money in High-Inflation Countries." Mimeographed. London School Econ., 1976.

———. "The Monetary Dynamics of Hyperinflation: A Reply." *J. Monetary Econ.* 3 (February 1977): 119-21.

Koot, R. S. "Nonconstant Coefficients of Expectation and the Recent Demand for Money." *J. Monetary Econ.* 1 (January 1975): 375-82.

Mussa, M. "Sticky Prices and Disequilibrium Adjustment in a Rational Model of the Inflationary Process." Mimeographed. London School Econ., 1976.

Sargent, Thomas J., and Wallace, Neil. "Rational Expectations and the Dynamics of Hyperinflation." *Internat. Econ. Rev.* 14 (June 1973): 328-50.

Silveira, A. M. "The Demand for Money: The Evidence from the Brazilian Economy." *J. Money, Credit and Banking* 5 (February 1973): 113-40.

On the Structure and Stability of Political Markets

W. Mark Crain

Virginia Polytechnic Institute and State University

This paper considers an organizational aspect of the market in which votes are exchanged for public-policy outcomes. Specifically, the effect on the stability and behavior of politicians of assigning the demanders of political products (i.e., voters) to geographic areas is addressed. Implications concerning the locational division of the "buyers" of political outcomes for collusive efforts by existing officeholders to restrict entry are drawn and tested empirically. The results indicate in effect that the institutional structure of political markets is an important aspect of the degree of rivalry among existing politicians and hence the extent of entry by nonincumbent candidates.

> . . . Machine politicians are simply the response to the fact that the electoral mass is incapable of action other than a stampede, and they constitute an attempt to regulate political competition exactly similar to the corresponding practices of a trade association. [Joseph Schumpeter][1]
>
> . . . The conditions appropriate to the assignment of customers will exist in certain industries, and in particular geographical division of the market has often been employed. [George Stigler][2]

I wish to thank James Buchanan, Thomas Deaton, Harold Elder, Thomas Saving, and Gordon Tullock for helpful suggestions and encouragement as this paper progressed. The critical comments offered by Roger Sherman, George Stigler, and Robert Tollison warrant special acknowledgment. I am responsible, of course, for any shortcomings that remain.

[1] Schumpeter 1962, p. 283.

[2] Stigler 1964, p. 48.

[*Journal of Political Economy*, 1977, vol. 85, no. 4]
© 1977 by The University of Chicago. All rights reserved.

I. Introduction

The policy outcomes supplied through the public sector are demanded ex ante by individuals with their votes. No legal recourse exists for fraudulent product advertising or even for failure of product delivery. Presumably, though no "meeting of minds" has occurred, exchange has.

This paper considers an organizational aspect of the market in which votes are exchanged for public-policy outcomes. Specifically, the effect on the stability and behavior of the "sellers" of assigning the "buyers" of political products to geographic areas is addressed. In Section II a conceptual framework for approaching the transactional arrangements between politicians and voters is presented. Section III draws some observable implications concerning the effects of geographical division of political representation and provides some empirical testing of these implications. Finally, Section IV offers concluding comments.

II. The Functioning of Fixed Political-Market Shares

The rivalry among candidates to obtain (temporary) property rights to operate the production apparatus of the public sector will, in the ideal, serve to protect the sovereignty of individual voter-consumers.[3] The degree of electoral competition among political entrepreneurs will of course vary and as such will result in departures from the ideal. Limitations on this periodic rivalry, then, can be expected to effect the desired functioning of the vote public policy exchange market.[4]

There are numerous reasons to suspect that the existing "firms" (i.e., incumbent officeholders) in the public-policy industry have incentives to restrict electoral competition. For example, the ability to extract rents from those seeking favorable regulatory legislation would be a function of a politician's expected tenure in office. In addition, the politician's influence over legislative outcomes is generally related to his length of service (because of seniority rules) and hence his ability to back promised political benefits to specific interest groups.[5] However, in a more immediate sense, each incumbent candidate for reelection to public office has an incentive to restrict electoral competition due to the presence of scale economies. The scale advantage present at the *individual*-election level is the majority voting rule which provides that the political entrepreneur who obtains half the customers can drive the other entrepreneur (or entrepreneurs) out of the market.[6] That is, while at an aggregated level elections may not be "won or lost," at an individual level they are.[7] Thus each incumbent entrepreneur has incentives to inhibit electoral rivalry for votes, both from within the field of existing firms and from potential (i.e., nonincumbent) entrants.

---

[3] For two early discussions of the "ideal" democratic structure, see Schumpeter (1962) and Becker (1958). An important implication of their definitional conception is that "the ultimate aim of each political party may be to acquire political power, but in equilibrium no one, including those 'in power,' has any political power" (Becker 1958, p. 107).

[4] The role played by rivalry in the mitigation of monopoly power in the organization of private economic activity has been discussed by Demsetz (1968, 1971), Telser (1969, 1971), and Crain and Ekelund (1976).

[5] For an empirical investigation of the relationship between the seniority power of congressional delegations to the U.S. House and influence over policy outcomes, see Crain and Tollison (1977) and the related study by Stigler (1976).

The rivalry among candidates to obtain (temporary) property rights to operate the representative apparatus takes place within geographic boundaries. The extent of these boundaries establishes the market area in which the exchange between voters and political candidates occurs. In other words, the periodic renewal of political contracts has a spatial element, just as in the case of exchange in private markets. Moreover, again as in the case of the latter, we might expect the degree of electoral competition to be affected by this spatial dimension of exchange.

The locational division of the "buyers" of political outcomes serves a two-fold function in the public-policy industry that may mutually benefit the existing "suppliers." First, it reduces considerably the gains from competition among the existing politicians. That is, the geographical assignment of voters to districts where residency requirements are in effect insures that no political entrepreneur can attract "customers" away from other incumbent politicians. Thus the competition for votes typically will not involve a rivalry between established firms but, rather, between an established firm and a potential (i.e., nonincumbent) entrant. The importance of this function of geographic representation, then, is that it virtually prevents an existing political entrepreneur from going into an election contest without an incumbent's advantage. Without such prevention the precampaign advantage of an incumbent would be muted should he have to face another incumbent also seeking reelection.[8] This first aspect is quite analogous to collusive practices in the organization of economic activity, where "fixing market shares is probably the most effective of all means of combating secret price reductions" among the existing firms (Stigler 1964, p. 44).

---

[6] This is essentially the argument made by Tullock (1965); see esp. pp. 464–65.

[7] Stigler (1972) has stressed that at an aggregated level (e.g., all the seats in a legislative assembly) elections may not be "won or lost" in the sense that the control of a certain percentage of the total number of seats is important even if it is less than a majority. Stigler argues, in effect, that such outcomes can be characterized as more usefully varying continuously because the control of fewer than 50 percent of legislative seats does not exclude an entire coalition or political party (as a group) from influencing public-policy outcomes. For some empirical verification of this argument, see Palda (1975) and Crain and Tollison (1976). The important point to note here is the distinction between election outcomes at the "industry" or aggregated level versus the "firm" or individual level.

[8] For some empirical estimates of the preelection advantage attributable to "incumbency," see Welch (1974, 1976) and Jacobson (1976). It should be noted that some incentives would still be present for incumbent politicians to engage in indirect election rivalry. In addition to interparty conflicts, there is some evidence that the competition for campaign funds is not limited to the "home district"; see Dawson and Zinser (1976). The point to be stressed with regard to the present argument is that legally segmented voting districts eliminate direct competition among incumbent candidates for the same voters.

A second function served by the geographical segmentation of political markets concerns its role in the establishment of barriers to potential entrants. Since incumbents are not likely to be threatened directly in the vote market by other incumbents (as discussed above), incentive to prevent an existing officeholder from enhancing his preelection advantage is reduced. That is, internal resistance to an existing entrepreneur using the powers of incumbency to strengthen his electability is slight, because the other officeholders in the industry have the same incentive. For example, to the extent that an incumbent's popularity is increased by his productivity in passing legislative outcomes desirable to a majority of his constituents, his logrolling feats will be enhanced by the fact that other incumbents stand to gain from vote trading for the same reason. In another light, all incumbents are likely to keep the rules of the game in their favor since they only run against nonincumbents.

Legal or constitutional restrictions are generally flexible enough to facilitate collusive measures on the part of incumbents (e.g., the passage of legislation providing expense accounts for travel to the home district or for mailing privileges, direct access to media coverage, influence over appointed positions) which can affect positively their brand-name capital and ability to command electoral support.[9] Even in a flexible legal setting, such collusive activity would surely tend to falter if incumbents were subjected to direct rivalry in the vote market. Given the possibility of direct competition, some incentives would be present to hinder the preelection advantage of an incumbent who addresses vote markets that overlap with other incumbents. In effect, incumbents would restrain the subsidized campaign activity of their colleagues, which could reduce further the costs (including information costs) of voting. In sum, the second function served by the locational division of vote markets is to reduce incentive to inhibit use of incumbency powers to establish or strengthen barriers to potential nonincumbent entrants.

The structure of political markets (e.g., strict geographic representation vs. at-large representation) has perhaps a more profound influence on the conduct of incumbent politicians in establishing entry barriers than do legal or constitutional regulations. In fact many of the latter might be deemed unnecessary under political settings conducive to direct incumbent competition in the vote market.

This discussion has stressed the gains from cooperative behavior among the entire class of incumbent politicians in restricting electoral competition.

---

[9] Recently several writers have argued convincingly that the passage of laws limiting the contributions and expenditures of campaign funds are nothing but incumbent protection acts. See esp. Kazman (1976) and Jacobson (1976). For an additional discussion of barriers to entry in politics, see Tullock (1965).

To the extent that affiliations with political parties are for a similar purpose (i.e., to coordinate the reelection of incumbent members of the party), it is still in the interest of incumbent officeholders of opposing parties to support agreements which are mutually beneficial in this respect.[10] As mentioned above (n. 9), limitations on campaign expenditures may provide an example of such an agreement.

Willingness to collude across party lines, however, is a function of the particular institutional determinants regarding the relative differential in the political power wielded by an officeholder who is a member of the majority party (e.g., committee chairmanships, influence over selecting positions in the executive branch, etc.). That is, at an individual-choice level, incentive to form protective agreements with incumbent members of another party will decline, the greater are the potential gains (or losses) in altering the party proportions of the legislature. In other words, the relevant trade-off facing an incumbent would be to seek to restrict electoral rivalry everywhere (i.e., to protect his own "seat" as well as those of the other incumbents) versus risking the entry of nonincumbent candidates in hopes of obtaining incremental proportions for his own party and thus incremental personal influence over legislative outcomes. Of course the "minimax" decision of a risk-averse incumbent would be to collude with other incumbents—regardless of his party's proportion—which suggests that reversals in the majority party controlling the legislative assembly would not be frequent. Party loyalty among the existing entrepreneurs in the public-policy industry, in this perspective, is dependent on individual wealth-maximizing behavior and indeed may be more apparent than real.

III. Implications and Empirical Findings

Several observable implications follow directly from the operational functions of geographic representation discussed in the preceding section. This section develops some of these implications into falsifiable hypotheses and offers some empirical testing. Implications regarding the impact of specific political-market structures are difficult to test at the federal level, since all direct vote competition for the same office (e.g., the Senate or the House of Representatives) generally occurs on a strictly geographic basis. Though the two U.S. senators from each state do confront a common geographic vote market, they stagger their confrontations temporally. At the state level, however, a fruitful empirical setting is provided, since different institutional arrangements concerning representative apportionment are in effect across the different states. Hence in this latter setting comparisons can be made across states to isolate different effects where alternative methods of apportionment are employed.

---

[10] A dictum by the late Sam Rayburn is illustrative of this role of political parties: "All incumbent Democrats who are in trouble get maximum assistance first. If anything is left over it will then be parcelled out to the best advantage" (quoted in Redding 1958, p. 303). Indeed the party apparatus may be more important in restricting nonincumbent entry at the level of primary elections than in general elections.

Before presenting the hypotheses, perhaps it is useful to clarify briefly two methodological points. First, the objective is to detect observable consequences attributable to the alternative institutional arrangements as they exist across states. This is in contrast to the separate, albeit interesting, issue of explaining the existing pattern of apportionment institutions. In other words, no theory is offered to explain the underlying differences in districting procedures as they have developed historically. Rather, the approach here is to take the set of institutional rules as given and examine outcomes of these differences. Perhaps the best analogy is to those studies which examine pricing or output behavior in different types of private-market structures, as opposed to studies which seek to explain why some industries tend to be competitively organized while others tend to be monopolized.

The second point concerns the nature of the dependent variable employed in this section. The relative stability of incumbent politicians, as measured by turnover, is examined in various state legislatures in order to obtain evidence on the degree of rivalry among candidates. That is, more stability implies less turnover and thus lower levels of entry by new candidates. While one might argue that high levels of new entry into politics will tend to result in a closer correspondence between ex ante voter preferences and ex post political outcomes, alternative arguments surely could be made to support the benefits of less frequent turnover.[11] This is to say that the effect in principle of free entry is not so clear in a political setting. There are some clear advantages to retaining existing political entrepreneurs because of such factors as start-up costs and on-the-job-training costs and some evidence about the past (versus promised) performance of a candidate.

The electoral process, moreover, can be viewed as an instrument to induce officeholders to behave in accordance with constituency preferences, and thus an increased probability of nonreelection would tend to weaken this effect. That is, in the extreme, if no incumbent were ever reelected, his behavior in the current term would be utterly irresponsible.[12] Hence entry of new entrepreneurs in a political setting (necessarily implying a turnover of existing suppliers) is not good per se. Frequent (involuntary) turnover does provide an indication, however, of the height of the entry barriers and the degree to which they are surmountable and

[11] For some discussion of this nature, see Rosenthal (1974), esp. pp. 166–85. Similarly, Crane and Watts contend that the high rate of turnover on standing committees which results from a general legislative instability "reduces the level of performance of both the lawmaking and watchdog functions of a representative body" (1968, p. 49).

[12] For a further theoretical discussion of the impact of succession rights on the behavior of officeholders, see Barro (1973).

therefore offers some evidence as to the overall extent of rivalry among candidates. That is, a high rate of turnover suggests that those incumbents who remain in office are likely to do so more because of closeness to voter preferences than because of institutional restrictions on voter alternatives.

We now turn to the presentation of each of the specific hypotheses to be tested empirically.

One fundamental difference in the apportionment procedures across states is the particular agency delegated authority to establish the representative districts. In 1974, for example, six different apportioning agencies held this legal authority across states. Among these agencies were the incumbent members of state legislatures.[13] If the incumbent members of state legislatures have incentives to retain their current property rights, we would expect that when given the power of apportionment they would employ this authority accordingly. Thus one broad implication is that we should observe differences between the stability of the legislative houses in those states allowing the legislature to apportion and in those opting for other apportionment agencies.

A second noticeable difference in apportionment practices across states is the use of multimember districting. This aspect of state representative government renders a rather straightforward implication concerning the specific functioning of fixed political-market shares. In effect, some states employ apportionment rules which place more than one representative in a single district, and in such cases incumbents face incumbents in reelection rivalry. This is in contrast to the arrangements in other states which practice strict geographic representation, that is, one representative per district. Thus a second implication is that we should observe differences between the stability of incumbents in those states where the segmentation of voters does not permit "cheating" or trying to attract voters away from other incumbents (i.e., in states employing strict geographic representatives) and in those states where incumbent vote markets overlap (i.e., in those states employing multimember districting).[14]

A third implication to be addressed empirically is the relationship between the number of voters ("potential customers") and the extent of electoral rivalry among political candidates. To draw once more on an analogy from the organization of economic activity, we would expect rivalry for votes (in those institutional settings where this is possible) to be more intense as the number of potential voters per representative decreases. This is because the potential gains from "cheating" or attracting votes away from another incumbent are offset by the potential losses in votes to other incumbents who similarly seek new customers. Moreover, as size of the vote market increases, a given level of cheating activity will achieve smaller percentage increases in vote shares. Thus, where the number of voters in a district is relatively large, we would expect less incentive to "cheat" or to engage in vote rivalry among existing political entrepreneurs.[15]

---

[13] The six categories of agencies which were responsible for state legislative apportionments in 1974 were boards of commission, federal courts, state courts, governors, legislatures, and secretaries of state (Council of State Governments 1976).

[14] Incumbents in the latter type of apportionment setting would still have the same incentives to agree to restrain direct vote rivalry, but in the absence of districting barriers such agreements would be harder to sustain due to the individual gains from cheating. Residency requirements might be viewed more accurately as simply raising the costs of vote rivalry. That is, it is conceivable that a candidate might pay the relocation costs of citizens in exchange for promises of future votes. However, the chances of a candidate being double-crossed are high due to ballot secrecy, thus reducing the expected benefits from incurring such transactional costs. This does, however, raise a further implication (which is not tested in the present work) that, where residency requirements for voting are less stringent, geographically divided markets would be less effective in prohibiting candidate rivalry.

The fourth and final implication to be investigated in the present paper concerns the influence of political parties on incumbent rivalry. Following the discussion in the previous section, the extent of cooperative effort among all incumbents to restrict entry will depend on the subjective evaluation of each existing officeholder as to the relative gains from altering party proportions versus enhancing the overall stability of the existing seats. The expected benefits from altering party proportions would be a function of the probability that enough candidates are elected from the same party as well as the particular institutional setting which I discussed previously. Thus the preelection closeness of party percentages might be expected to affect this underlying decision calculus. That is, given a level of benefits attributable to being an incumbent member of the majority party, the fewer seats that are necessary to affect this majority, the higher are the expected returns to rivalry. The logic here in effect is that less collusive activity will occur (implying more entry) as the preelection party proportions approach equality.

The empirical model employed to examine these four implications across states is specified thus:

$$\begin{aligned} N_i = \beta_0 + \beta_1 A_i + \beta_2 MD_i + \beta_3 ID_i + \beta_4 D_i \\ + \beta_5 V_i + \beta_6 S_i + \beta_7 DP_i + \beta_8 EP_i + \mu_i \end{aligned} \tag{1}$$

where

N_i = the number of nonincumbent candidates elected to the legislative body in state i in 1974[16]

---

[15] For a discussion of this point in the context of industrial organization, see Stigler (1964), esp. p. 49.

[16] If the size of a legislative assembly increased in this election, the election of a nonincumbent candidate would not necessarily imply the defeat of an incumbent. There were three cases in 1974 in which the size of legislatures changed, but in two of these instances the number of seats was reduced. The size of the lower houses in Alabama and Maryland was reduced by one seat each to 105 and 141, respectively. The size of the upper house in Maryland, however, was expanded by four seats to 47. Thus Maryland is excluded from the regression sample for the upper houses.

A_i = a dummy variable denoting the apportioning agency in state i, where $A_i = 0$ if the legislature has this power and $A_i = 1$ otherwise
D_i = a dummy variable denoting the presence of multimember districting in the legislative body in 1974, where $D_i = 1$ if state i has *any* multimember districts and $D_i = 0$ otherwise
M_i = the number of multimember districts in the legislative body of state i in 1974
MD_i = an interaction term, the product of M_i and D_i
I_i = the average number of incumbents per district running for reelection in state i in 1974
ID_i = an interaction term, the product of I_i and D_i
V_i = the average size of a single legislative district in state i in 1974
S_i = the total number of seats in the legislative body of state i that were up for election in 1974
P_i = the ratio of the number of seats held by the minority party to those held by the majority party in the legislative body of state i prior to the 1974 elections
DP_i = an interaction term, the product of P_i and D_i, which differs from zero in those states *with* multimember districts
EP_i = an interaction term which differs from zero in those states *without* multimember districts (i.e., the product of P_i and E_i where $E_i = 1$ in states without multimember districts and $E_i = 0$ elsewhere)
μ_i = a random-disturbance term

The eight coefficients defined in equation (1) are estimated from 1974 data on the variables across states in the United States (Council of State Governments 1976). The upper and lower houses of the state legislative assemblies are examined separately since the relevant incentives on the part of incumbent officeholders are operative primarily with regard to competition for a specific type of office. The results for state lower houses are presented in equation (2) and for state upper houses in equation (3):[17]

$$\begin{array}{lllllll}
N_i^L = & 11.88 & +\ 7.85A_i & +\ 0.37MD_i & +\ 13.11ID_i & -\ 46.67D_i & \\
 & (1.52) & (1.81) & (1.98) & (2.62) & (-2.58) & \\
 & -\ 0.62 \times 10^{-4}V_i & & +\ 0.29S_i & -\ 6.95DP_i & -\ 13.39EP_i & (2) \\
 & (-1.67) & & (6.43) & (-0.62) & (-1.37) &
\end{array}$$

$$R^2 = .7828, \qquad F_{(8,39)} = 17.57, \qquad N \text{ observed} = 48$$

---

[17] Louisiana and Mississippi were omitted from eq. (2) because no elections were held in the lower houses of these states in 1974. Likewise, six states did not hold elections for

$$
\begin{aligned}
N_i^U = {} & \underset{(-0.25)}{-0.72} + \underset{(3.27)}{5.59A_i} + \underset{(1.63)}{0.70MD_i} + \underset{(0.72)}{2.38ID_i} - \underset{(-1.16)}{16.66D_i} \\
& \underset{(-1.43)}{-\ 0.10} \times 10^{-4}V_i + \underset{(6.06)}{0.38S_i} + \underset{(1.25)}{9.81DP_i} + \underset{(0.03)}{0.10EP_i} \qquad (3)
\end{aligned}
$$

$$R^2 = .5943, \qquad F_{(8,34)} = 11.19, \qquad N \text{ observed} = 43$$

The results for each independent variable are discussed and interpreted in turn.

The estimated coefficient for the variable denoting the agency responsible for legislative apportionment, A_i, is positive for both the state lower and upper houses, occurring at a .05 level of significance in the former case and at a .005 level in the latter. The positive sign of this coefficient indicates that the level of nonincumbent entry is lower in those states where the incumbent members of the respective legislative body are delegated apportionment rights. The magnitude of this effect is greater in the case of state lower houses, where the power to apportion (on average) saves the political lives of over seven incumbents per election; while in the case of state upper houses apportionment rights in the hands of incumbents, ceteris paribus, reduces the average turnover per election by about five seats. This initial finding supports the specific contention that, when endowed with legal authority to establish legislative boundaries, incumbents can effectively do so in accordance with their self-interest, for example, to secure an extended tenure in office.

The extent of multimember districting can be varied in two ways, both of which have been accounted for in the empirical specification in equation (1). One, states can vary the number of districts which are to be served by more than one representative; or two, they can vary the number of representatives per district.[18] The variable accounting for the former method in equation (1) is MD_i, and for the latter, ID_i. The MD_i and ID_i are interaction variables with D_i, while D_i is included independently in equation (1) as strictly a shift parameter. The independent effect estimated for the number of multimember districts, MD_i, is positive in both equations above. For the lower houses (eq. [2]), the estimated coefficient is significant at the .05 level; and for upper houses (eq. [3]) it is significant at the .10 level. These two estimated results suggest that the entry of new politicians is affected marginally by the number of multimember districts in the legislative body. More specifically, this result can be interpreted to imply that the more incumbents placed into direct vote

---

upper-house members in 1974. The states omitted from eq. (3) for this reason were Kansas, Louisiana, Minnesota, Mississippi, South Carolina, and Virginia. In addition, Maryland was omitted from eq. (3) for the reason cited in n. 16 above.

[18] For some further discussion and empirical investigation of institutional factors as variables in the political structure of state governments, see Crain (1976).

rivalry, the easier it is for nonincumbent candidates to enter. As discussed in Section II above, the logic here is that collusive agreements to restrain electoral competition will be harder to maintain because of the potential gains in votes, and the ability of incumbents to construct entry barriers will be inhibited.[19]

Similarly, as the number of incumbents per multimember district, ID_i, increases, the entry of new members into state lower houses increases. The estimated coefficient for this effect in state lower houses is positive and significant at the .01 level. This variable does not appear at a statistically significant level for state upper houses (eq. [3]), which is perhaps due to its rather narrow range in the case of the latter. The sign is, however, also positive. Again, this variable is interpreted here to imply that the increased difficulty of incumbent collusion due to larger numbers per district will result in relatively lower entry barriers and hence the observed pattern of more new entrants.

The coefficient estimated for V_i indicates the relationship between average size of the districts in the legislative body and the rate at which new political entrepreneurs entered in the 1974 elections. The effect estimated in both equations is negative and significant at the .10 level for both types of state legislative assemblies. As suggested earlier in this section, this negative relationship is precisely what we would expect, by analogy to the organization of economic activity. That is, as the total size of the vote market increases, we would expect the gains from "cheating" (i.e., seeking to pull voters away from the support base of other incumbent representatives) to be offset by the potential losses from other candidates behaving likewise. As the average size of the constituencies increases, we would expect less incumbent rivalry and thus the observed result of fewer new entrants. Moreover, the shift parameter, D_i, which controls for differences in the two types of districting procedures, is negative and significant in the equation for state lower houses. This negative all-or-none shift indicates that the size of the vote market affects a smaller "base" level of entry in those lower houses with multimember districts.

As one would suspect, the rate of entry by new politicians is positively and significantly related to the number of seats up for election. This variable is included in the specification of equation (1) primarily to control for the absolute size of legislative bodies in examining the other somewhat more interesting institutional variables concerning the structure of political markets. However, one would expect the extent of collusive behavior to diminish as group size increases.

---

[19] Rivalry among incumbent politicians also has direct external benefits to nonincumbents seeking election in such situations, because the impact of one incumbent challenging the credibility and past performance of another would carry a larger impact than identical attacks by a nonincumbent challenger. That is, the advantages from brand-name capital and accessibility to media coverage mentioned above can be employed "positively" or "negatively," with roughly similar results. Further, incumbent competition could tend to reduce generally the costs to voters of acquiring information about the past performance of officeholders seeking reelection. Such information would be generated during the rivalry because each incumbent potentially could expand his market share of votes by exposing the foibles of his colleagues. This effect is in addition to the indirect effects discussed in this paper concerning the collective ability of incumbents to create or heighten entry barriers to nonincumbent challengers. These barriers should be generally lower in those settings where incumbents share common vote markets.

Finally, the results estimated for the effects of the preelection party proportions on nonincumbent entry are somewhat mixed. The estimated coefficient for DP_i (which is nonzero for those legislatures where multimember districting is employed) is positive and significant at the .15 level for state upper houses but does not differ significantly from zero when estimated for state lower houses. The estimated coefficient for EP_i (the variable which is nonzero in states employing single-member districting) is negative and significant at the .10 level for the lower legislative bodies, but this estimate for the upper houses does not differ significantly from zero.

Thus, following the logic of the decision calculus posited above, in state upper houses we do observe a detectable tendency for incumbents to opt for potential gains from altering party proportions. Only in this case do we find consistent increases in the level of nonincumbent entry as the preelection party proportions approach equality. This suggests that differences in the value of incumbency issuing from institutional aspects of majority control are more pronounced in the case of state upper houses. However, this appears to hold only in state upper houses that employ multimember districting, perhaps because the opportunity cost of partisan rivalry is higher in single-member districting states. That is, if the structural advantages to incumbent cooperation are greater under strict geographic representation, we would expect less partisan rivalry when incumbents calculate the relevant trade-offs in this latter type of setting.

These results concerning the role of political parties in state legislatures indicate that partisan loyalties in general are not sufficient to generate electoral competition among the existing set of incumbents. The institutional structure of vote markets seems to influence party aspects of candidate behavior and, in a broader sense, party control over representatives. While the empirical findings on the role of parties are admittedly tentative, this economic formulation of the choice problem facing incumbent officeholders may provide a useful framework for further investigation.

IV. Concluding Comments

This paper has relied heavily on theories from the organization of economic activity in developing a conceptual framework for viewing the transactional arrangements between politicians and voters. Specifically, I examined the relationships between some aspects of representative apportionment and the conduct of the suppliers of political products. Implications regarding the effects of alternative apportionment institutions on the degree of rivalry among incumbent political entrepreneurs were tested in the context of state governments. The empirical findings supported the contention that the structure of political markets influences the stability of existing members of state legislative bodies and hence the entry of new members.

The role of political parties on electoral competition was examined in the context of a simple wealth-maximizing choice problem. The results in this case indicated that the expected benefits from partisan rivalry were also dependent upon the particular institutional structures employed in state legislative assemblies. In most settings, the gains from party rivalry did not appear to offset the gains from cooperative effort among all incumbents.

The level of new entry has further implications concerning the correspondence between voter preferences and outcomes in the public policy industry. In addition, the use by over half the states of political-market structures which function to foster incumbent rivalry may provide some clues as to the vast disparity which persists between the rate of non-incumbent entry into politics at the state level as opposed to the federal level.[20]

References

Barro, Robert J. "The Control of Politicians: An Economic Model." *Public Choice* 14 (Spring 1973): 19–42.

Becker, Gary S. "Competition and Democracy." *J. Law and Econ.* 1 (October 1958): 105–9.

Council of State Governments. *Book of the States, 1976–1977.* Lexington, Ky.: Iron Works Pike, 1976.

Crain, W. Mark. "Institutional Factors and the Production of Political Products." Mimeographed. Virginia Polytech. Inst. and State Univ., 1976.

Crain, W. Mark, and Ekelund, R. B. "Chadwick and Demsetz on Competition and Regulation." *J. Law and Econ.* 19 (April 1976): 149–62.

Crain, W. Mark, and Tollison, R. D. "Campaign Expenditures and Political Competition." *J. Law and Econ.* 19 (April 1976): 177–88.

———. "The Influence of Representation on Public Policy." *J. Legal Studies* (1977), in press.

Crane, W. W., and Watts, M. W. *State Legislative Systems.* Englewood Cliffs, N.J.: Prentice-Hall, 1968.

Dawson, P. A., and Zinser, J. E. "Political Finance and Participation in Congressional Elections." *Annals American Acad. Polit. and Soc. Sci.* 425 (May 1976): 59–73.

---

[20] For an extensive overview of the tenure patterns in the U.S. House of Representatives, see Polsby (1968).

Demsetz, Harold. "Why Regulate Utilities?" *J. Law and Econ.* 11 (April 1968): 55–65.

———. "On the Regulation of Industry: A Reply." *J.P.E.* 79, no. 2 (March 1971): 356–63.

Jacobson, G. "Practical Consequences of Campaign Finance Reform: An Incumbent Protection Act?" *Public Policy* 24 (Winter 1976): 1–32.

Kazman, S. "The Economics of the 1974 Federal Election Campaign Act Amendments." *Buffalo Law Rev.* 25, no. 2 (1976): 519–43.

Palda, K. S. "The Effect of Expenditure on Political Success." *J. Law and Econ.* 28 (December 1975): 745–71.

Polsby, N. W. "The Institutionalization of the U.S. House of Representatives." *American Polit. Sci. Rev.* 62, no. 144 (1968): 144–68.

Redding, J. *Inside the Democratic Party.* Indianapolis: Bobbs-Merrill, 1958.

Rosenthal, A. *Legislative Performance in the States.* New York: Free Press, 1974.

Schumpeter, J. A. *Capitalism, Socialism, and Democracy.* 3d ed. New York: Harper & Row, 1962.

Stigler, George J. "A Theory of Oligopoly." *J.P.E.* 72, no. 1 (February 1964): 44–61.

———. "Economic Competition and Political Competition." *Public Choice* 13 (Fall 1972): 91–106.

———. "The Sizes of Legislatures." *J. Legal Studies* 5 (January 1976): 17–34.

Telser, Lester G. "On the Regulation of Industry: A Note." *J.P.E.* 77, no. 6 (November/December 1969): 937–52.

———. "On the Regulation of Industry: Rejoinder." *J.P.E.* 79, no. 2 (March/April 1971): 364–65.

Tullock, Gordon. "Entry Barriers into Politics." *A.E.R.* 55 (May 1965): 458–66.

Welch, W. P. "The Economics of Campaign Funds." *Public Choice* 20 (Winter 1974): 83–97.

———. "The Effectiveness of Expenditures in State Legislative Races." *American Polit. Q.* 4 (July 1976): 333–56.

On Uncertain Lifetimes

Robert J. Barro and James W. Friedman

University of Rochester

This paper contrasts consumer choice under uncertain lifetimes with the behavior that would arise if each individual's lifetime were announced at birth. In a model that includes life insurance and excludes investments in human capital, the expected utility under uncertain lifetimes exceeds that under known lifetimes when the latter expectation is based on preannouncement survival probabilities. This conclusion emerges, first, because the model without human capital contains no planning benefits from knowledge of the horizon and, second, because the prior announcement of lifetimes forces risk-averse consumers to undertake an extra gamble that they could otherwise avoid by using life insurance.

This paper explores the role of uncertain lifetimes in an intertemporal model of consumer choice. The analysis compares the behavior of consumption and utility under a stochastic horizon with that which would arise under certainty where each individual's length of life is announced at birth. In a model that includes life insurance and excludes investments in human capital, the expected utility under uncertain lifetimes exceeds the expected utility under known lifetimes when the latter expectation is based on preannouncement (ex ante) survival probabilities. This conclusion emerges, first, because the model without human capital accumulation contains no planning benefits from knowledge of the horizon; and, second, because the prior announcement of the time of death forces risk-averse consumers to undertake an extra gamble that could have been avoided by the use of life insurance in the uncertainty case.

The possibility of accumulating human capital implies planning benefits from knowledge of the time of death. The ability to match investments with the horizon implies that expected lifetime wealth is lower in the case of uncertain lifetime than in the case where lifetimes are

[*Journal of Political Economy*, 1977, vol. 85, no. 4]
© 1977 by The University of Chicago. All rights reserved.

announced at birth. The effects on expected lifetime utility are ambiguous because the planning benefit from knowledge of the horizon offsets the cost associated with risk aversion.

One other result is that, contrary to some earlier suggestions, an uncertain lifetime is not a rationale for time preference when "perfect" life insurance is admitted. Although survival probabilities below unity imply a discount on future utility, they imply a corresponding discount in the budget constraint on future income and expenditure. On net, there is no effect of uncertain lifetime on the preference for current consumption relative to future consumption.[1]

In Section I a model is examined in which the individual's lifetime is uncertain and there is no human capital accumulation. Section II deals with a model in which the lifetime is known, though randomly determined at the start. In Section III the models of the earlier sections are modified to allow for the accumulation of human capital.

I. Uncertain Lifetime with No Human Capital

Consider a consumer who lives no more than T periods, with the probability of dying at the end of the ith period being p_i. It is assumed that, at the end of the first period, a random draw occurs. With probability p_1 the consumer dies. In general, if the consumer lives during period i, a random drawing with a death probability of $p_i/\sum_{j=i}^{T} p_j$ takes place. It is assumed that $p_i \geq 0$ for $i = 1, \ldots, T - 1, p_T > 0$ and $\sum_{i=1}^{T} p_i = 1$. The consumer is viewed as maximizing expected lifetime utility, which is given by

$$\bar{U} = u(c_1) + (1 - p_1)u(c_2) + \cdots + (1 - p_1 - \cdots - p_{T-1})u(c_T), \tag{1}$$

where c_i is consumption and $u(c_i)$ is the satisfaction enjoyed during the ith period of the lifetime. Let $q_i = 1 - \sum_{j=1}^{i-1} p_j (i > 1)$ be the probability as of birth that the person survives into the ith period. Then, with $q_1 = 1$, equation (1) can be written as

$$\bar{U} = \sum_{i=1}^{T} q_i u(c_i). \tag{2}$$

The function, $u(c)$, is assumed to be increasing, concave, and differentiable. Thus, $u' > 0$ and $u'' < 0$. Note that $u'' < 0$ is often interpreted as "risk aversion," and also that the lifetime utility function contains no discounting of future enjoyments beyond that which is included in the survival probabilities.[2]

---

[1] This result can also be seen in case C of Yaari (1965), which is a continuous time version of the discrete time model in Section I.

[2] Some comments are made on the effects of discounting in n. 5 of Section II.

Assume that the market interest rate is zero (see n. 5 below) and that the individual can purchase annuities and life insurance at actuarially fair prices, as well as being able to borrow and lend. These market opportunities make possible such diverse contracts as "negative annuities," an example of which is a contract where the consumer receives \$1.00 at the start of period 1 and repays $1/(1 - p_1)$ contingent on survival at the start of period 2. To duplicate this contract with a loan secured by life insurance, the consumer borrows $1/(1 - p_1)$ dollars at the beginning of period 1 and pays a premium of $p_1/(1 - p_1)$ dollars for a 1-period life insurance contract that pays off $1/(1 - p_1)$ dollars at the beginning of period 2 should he die by that time. For an additional discussion of contracts of this sort, see Yaari (1965). The present framework is a discrete version of his (continuous time) case *C*.

Taking the purchase of life insurance into account, the budget constraint is a requirement of equality between expected lifetime income and expected consumption expenditures. Let y be the income received during each period of life, A the initial assets of the consumer (which may be negative), and $\bar{W}$ the expected value of lifetime wealth (which must be positive). Then the budget constraint is

$$\sum_{i=1}^{T} q_i c_i = y \sum_{i=1}^{T} q_i + A = \bar{W}. \tag{3}$$

The term $\sum_{t=1}^{T} q_t = \bar{l}$ is the expected length of life; hence, expected lifetime wealth and expected length of life are related by

$$\bar{W} = y\bar{l} + A. \tag{4}$$

It is immediate that the maximization of $\bar{U}(c_1, \ldots, c_T) = \sum_{i=1}^{T} q_i u(c_i)$, subject to the budget constraint (eq. [3]), implies the first-order conditions,

$$u'(c_1) = u'(c_2) = \cdots = u'(c_T). \tag{5}$$

That these first-order conditions correspond to a (constrained) utility maximum follows from the concavity of $\bar{U}$ and the linearity of the budget constraint.

Thus, an optimal plan requires a constant level of consumption over time, $\bar{c} = \bar{W}/\bar{l} = y + A/\bar{l}$, and the expected value of attained lifetime utility is

$$\bar{U} = \bar{l}u(\bar{c}) = \bar{l}u(y + A/\bar{l}). \tag{6}$$

It has been argued that the uncertainty of survival is a rationale for the inclusion of a discount factor on future utility.[3] This observation suggests, in turn, that a person living in a world such as that modeled here would exhibit a positive rate of time preference—that is, with a zero market rate of interest, the optimal consumption plan would call for a rate of consumption that declines over time. This argument turns out to be invalid because the same "probability discount factors" that appear in the objective function, equation (2), appear also in the budget constraint, equation (3).

---

[3] See Bohm-Bawerk [1891] 1971, p. 255; Pigou 1962, p. 26; and Friedman 1969, p. 22.

II. Known Lifetime with No Human Capital

Consider modifying the model of Section I to allow for a length of life that is randomly selected at the start and announced at that time. As before, the probability of dying at the end of the ith period is p_i. However, at the moment of birth a random draw is assumed to occur, and the consumer's length of life, t, is determined and announced to him. It is also announced to any economic agent who might make an insurance contract with him. After the announcement of t, the consumer's objective function is

$$U = \sum_{i=1}^{t} u(c_i), \tag{7}$$

and his budget constraint is

$$\sum_{i=1}^{t} c_i = yt + A = W(t). \tag{8}$$

Note that the ex ante survival probabilities are absent from both the objective function and the budget constraint. Of course, the ex post survival probabilities are present. They are 1 for each of the first t periods and 0 thereafter.

Again, the optimal consumption plan calls for a constant level of consumption over time. For an optimum,

$$u'(c_1) = u'(c_2) = \cdots = u'(c_t) = u'(c^*), \tag{9}$$

where

$$c^* = y + A/t. \tag{10}$$

The attained level of lifetime utility is

$$U(t) = tu(c^*) = tu(y + A/t). \tag{11}$$

The interesting comparison to make with the model of Section I concerns a consumer who is at the beginning of his life and has not yet observed the outcome of the random draw that determines t, his length of life. His expected lifetime utility under the conditions of the present section is

$$E[U(t)] = E[tu(y + A/t)] = \sum_{t=1}^{T} p_t tu(y + A/t). \tag{12}$$

Meanwhile, the maximum value of expected lifetime utility that is attained in the model of Section I is

$$\bar{U} = \bar{l}u(y + A/\bar{l}) = U[E(t)]. \qquad (13)$$

Accordingly, the question of whether the consumer is better off under the risk of Section I or that of the present section comes down to a comparison between $E[U(t)]$ and $U[E(t)]$. If $U(t)$ is concave, then Jensen's inequality states that $E[U(t)] \leq U[E(t)]$.[4] It is easily verified that the concavity of $u(c)$ implies the concavity of $U(t)$; hence Jensen's inequality does hold. Given the conditions that $A \neq 0$, that u is strictly concave, and that at least two of the p_t are greater than zero, Jensen's inequality holds strictly. Under these circumstances, the regime of Section I is preferable in the sense that it yields the consumer a higher value of expected lifetime utility, as evaluated at the beginning of life. The key element that produces this result is that the consumer's ability to buy insurance gives him the same expected lifetime utility as he would have if his lifetime were known with certainty to be of length $\bar{l}$, the mean length of life.[5]

In the model of the present section, the consumer behaves optimally, making his decision after his length of life is determined. He does not escape having to take part in a gamble—namely, the gamble associated with the determination of t.[6] Under the other model, his ability to insure allows him to consume $y + A/\bar{l}$ in each period he lives, which has the effect of removing him from part of the gamble into which nature placed him. He does not escape the presence of uncertainty of length of life, but he does escape the utility cost associated with having to bear the consequent risk of an uncertain value of W/t—the per period amount of wealth.[7]

The two regimes yield the same utility if $A = 0$; however, the expected utility derived from the regime of Section I would again be higher if income, y, were not constant. The simplest case of nonconstant income occurs when the consumer is assumed to "work" for a number of years, having the same income per year, then to "retire" at some given age and have no employment income thereafter. Such assumptions, without consideration of a randomized length of life, occur in Modigliani and Brumberg (1954).

---

[4] On Jensen's inequality, see Mood, Graybill, and Boes (1974), p. 72.

[5] Since the time pattern of consumption is uniform under both known and unknown horizons, there are no "planning benefits" that derive from knowledge of the time of death. Introduction of an interest rate or of a subjective rate of discount on future utility would eliminate the constancy of consumption over time but would not eliminate the correspondence of the time pattern of consumption for the two cases. An example of planning benefits from known lifetimes that are associated with human capital accumulation is described in Section III.

[6] There is assumed to be no possibility of risk pooling before the announcement on lifetime occurs. Insurance possibilities are irrelevant after the announcement when the length of life is known.

[7] Hirshleifer (1971, p. 568) discusses some analogous situations in which information may not be socially useful. The length of life can be considered part of the individual's endowment in Hirshleifer's model.

In sum, the conclusion can be expressed as follows. Consider two regimes: regime 1 in which all individuals are born with a common probability density that determines the length of life; and regime 2 in which individuals are labeled at birth with a value of t that is drawn from the same distribution. The model implies that regime 1 is preferable—in the sense that, if both regimes were technically feasible, all (risk-averse) members of society would opt for regime 1.

The analysis is also suggestive for the optimal amount of investment to undertake in activities that yield information about individual life expectancy. The type of argument may provide a rationale for maintaining some ignorance about the state of health, especially where terminal illness may be involved.

III. Human Capital

The planning benefits from known lifetimes can be illustrated by introducing a variable amount of human capital. In a simple case all of the (nondepreciable) amount of human capital, K, is acquired during the first period of life. Suppose that income in each period is still constant over time but is now a function of the amount of human capital, as expressed by the function, $f(K)$, where $f' > 0$ and $f'' < 0$.

In the case of an uncertain horizon (the model of Section I), the only change is a modification of expected lifetime wealth (eq. [4]) to

$$\bar{W} = f(K)\bar{t} + A - K. \tag{14}$$

Choosing K to maximize expected lifetime utility (when an interior maximum exists) results in the first order condition,

$$f'(K) = 1/\bar{t}. \tag{15}$$

Letting $\hat{K}$ denote the optimal quantity of human capital, the optimal rate of consumption is

$$\hat{c} = f(\hat{K}) + (A - \hat{K})/\bar{t} = \hat{W}/\bar{t}, \tag{16}$$

and the expected value of lifetime utility is

$$\hat{U} = \bar{t}u(\hat{W}/\bar{t}) = \bar{t}u[f(\hat{K}) + (A - \hat{K})/\bar{t}]. \tag{17}$$

Turning now to the model in which the lifetime is determined before any actions are taken, lifetime wealth is

$$W(t) = f(K)t + A - K, \tag{18}$$

and the optimal value of human capital, $K(t)$, is the solution to

$$f'(K) = 1/t. \tag{19}$$

The assumptions on f' assure that equation (19) has at most one solution with $K > 0$. With a knowledge of the lifetime, individuals can match K with t, rather than basing the investment decision on $\bar{t}$. Optimal consumption and the maximized value of attained lifetime utility are

$$c(t) = W(t)/t = f[K(t)] + [A - K(t)]/t, \tag{20}$$

$$U(t) = tu[W(t)/t] = tu\{f[K(t)] + [A - K(t)]/t\}. \tag{21}$$

Consider, again, an ex ante calculation of expected utility when t is distributed in accordance with $(p_1, \ldots, p_T)$. It is straightforward to show that expected lifetime wealth is higher under the case of known horizon—that is,

$$E[W(t)] > W[E(t)]. \tag{22}$$

This conclusion follows from the convexity of W in t and the application of Jensen's inequality.[8] Intuitively, it may be thought of as resulting from the ability to make the investment decision after the lifetime is known, which provides a perfect match between K and t.

It is now not clear whether, in the presence of human capital investment (or analogous considerations that produce planning benefits from knowledge of the horizon), the consumer is better off having his length of life randomly determined prior to any decisions at the start of period 1, or having a random drawing at the end of each period that he survives. In the former situation he can make a perfect match of human capital investment to his (known) lifetime, but he cannot take out insurance that guarantees him a per period amount of consumption that corresponds to living the mean length of life. Obviously, which of these two elements predominates depends on the particular utility function, investment function, and probability distribution that apply.

References

Böhm-Bawerk, Eugen von. *The Positive Theory of Capital.* Translated by William Smart. Freeport, N.Y.: Books for Libraries, 1971.

Friedman, Milton. *The Optimum Quantity of Money.* Chicago: Aldine, 1969.

Hirshleifer, Jack. "The Private and Social Value of Information and the Reward to Inventive Activity." *A.E.R.* 61 (September 1971): 561–74.

Modigliani, Franco, and Brumberg, R. "Utility Analysis, and the Consumption Function: An Interpretation of Cross Section Data." In *Post Keynesian Economics,* edited by K. Kurihara. New Brunswick, N.J.: Rutgers Univ. Press, 1954.

Mood, A.; Graybill, F.; and Boes, D. *Introduction to the Theory of Statistics.* 3d ed. New York: McGraw-Hill, 1974.

Pigou, A. C. *The Economics of Welfare.* 4th ed. London: Macmillan, 1962.

Yaari, Menahem E. "Uncertain Lifetime, Life Insurance, and the Theory of the Consumer." *Rev. Econ. Studies* 32 (April 1965): 137–50.

---

[8] Convexity of $W(t)$ can be seen by differentiating eq. (18) and using eq. (19): $W''(t) = -(t^3 f'')^{-1} > 0$.

Consumer Horizon: Further Evidence

Khan A. Mohabbat and Evangelos O. Simos

Northern Illinois University

This paper provides new evidence on consumer subjective discount rate and the consumer horizon. Deriving the subjective interest rate for the United States, directly from total private wealth, the findings support Friedman's contention of a 3-year average horizon. Using the varying parameter method of estimation, a time series of factor of proportionality has been constructed and its behavior analyzed over the sample period 1929–69. Finally, the correlation between the permanent elements, permanent and transitory elements, and the transitory elements is verified. The results overwhelmingly support Friedman.

The length of the consumer horizon recently has been subjected to a great deal of debate, the outcome of which has important implications for the validity of the Permanent Income Hypothesis (PIH) as it rests on a consumer horizon of certain duration. For instance, if the length of horizon were short, say 1 year, there will be little distinction between the conclusions derived from the PIH and the Current Income Hypothesis (CIH) (Holbrook 1967, p. 754). On the other hand, if the length of the horizon were substantially longer than a year, the policy implications based on PIH will differ significantly from those of the CIH.

Friedman, himself, points out that a 3-year time horizon gives a good approximation of permanent income (1957, p. 150; 1963, p. 25). That is to say, given a 3-year horizon, any changes in income lasting fewer than 3 years in duration would be treated by the household as transitory, while one lasting 3 years or more would be considered as permanent. Laumas (1969), Landsberger (1971), Laumas and Mohabbat (1972) and Darby (1974) have provided varying degrees of support for

The authors, associate professor and graduate research assistant, respectively, wish to express their gratitude to Prem Laumas, James Gherity, and Edward Prescott for their helpful comments.

[*Journal of Political Economy*, 1977, vol. 85, no. 4]
© 1977 by The University of Chicago. All rights reserved.

Friedman's contention regarding the length of horizon.[1] However, Griliches et al. (1962), Holbrook (1967), and Wright (1969) provide evidence for a much shorter length of horizon.[2]

The wide disparity among the empirical findings of the above studies is due mainly to differences in the methodology and the definition of income and consumption that are employed to calculate the consumer discount rate. In this study we use a more direct approach to provide new evidence regarding the consumer horizon and the subjective discount rate, r.

Our method has several advantages over other studies. (1) We do not approximate permanent income as a weighted average of past and current income as is the case with other studies. As required by the PIH, we derive permanent income directly from total private wealth. Total wealth is derived from human and nonhuman resources. (2) Our consumption data are in strict accord with the PIH that includes expenditure on consumer nondurable goods and services of durable goods; that is, only a part of consumer spending on durable goods in any 1 year will be counted as that year's consumption, the balance constitutes consumer saving in the form of durable goods. Since durable goods last for several years their services will be added to consumption in the future periods. Other studies, due to lack of data on services of consumer durable goods, have employed either total consumer expenditures or expenditures on nondurables only. Neither definition meets the requirements of the PIH. (3) Most studies have either eliminated the war years or developed techniques to dampen the effects of those years. These attempts are, at best, rough approximations of the actual situation. It is our contention that PIH, if valid, should be able to explain consumer behavior during economic upheavals as well as economic stability. Thus our time-series data for the United States cover the period of 1929–69. (4) Our estimates of r, the subjective discount rate, are directly derived from total private wealth, thus avoiding the roundabout methods of estimation as employed by other studies.

This study proceeds as follows: Section I deals with methodology and empirical results and Section II presents concluding remarks. The data on income, consumption, and wealth for the United States, 1929–69, are obtained from Christensen and Jorgenson (1972, pp. 336, 337, 358).

---

[1] Landsberger's consumer discount rate based on Israeli data ranges between .17 and .45, yielding a corresponding horizon of 2.22–5.88 years. Laumas using Canadian data and Laumas and Mohabbat U.S. data have obtained identically the same horizon as Friedman. Although Darby, using U.S. data, derives a large number of consumer horizons, the one he prefers is 10.5 years in duration.

[2] Holbrook and Wright using U.S. time-series data obtain a consumer horizon of 7–13 months' and 1.25–1.5 years' duration, respectively. Griliches et al., also using U.S. data, estimate the horizon to fall between 3.6 and 8.4 months. It is essential to say that the data used in the case of Griliches et al. are quarterly data and have lagged consumption as an explanatory variable. Since they employ ordinary least-squares estimation procedure, the results concerning the horizon are downward biased.

I. Methodology and Empirical Results

Estimation of Consumer Discount Rate, r

The theoretical construct of PIH relates total wealth as the source of permanent income as

$$y_p = r \cdot W \tag{1}$$

where y_p is permanent income, W is total wealth consisting of the sum of human plus nonhuman wealth, and r is considered to be some form of subjective interest rate taking into account such things as the rates of returns on equities and bonds as well as rates applicable to human wealth and to physical goods.

In equation (1) only W is observable; y_p and r cannot be observed. Following Friedman (1957, pp. 21–22), (1) can be written as

$$Y_m - y_t = rW \quad \text{or} \quad y_m = rW + y_t, \tag{2}$$

with $\bar{y}_t = 0$ where y_m is measured income that is observable, y_t is transitory income which is not observable, and $\bar{y}_t$ is the average value of transitory income. With regard to y_t for cross-section data, Friedman (1957, pp. 21–22) states that "... the transitory component [of income] is to be interpreted as reflecting all ... factors that are likely to be treated by the unit affected as 'accidental' or 'chance' occurrences. ... For any considerable group of consumer units, the resulting transitory components tend to average out ... and the mean transitory component would be zero." The case for the average value of $y_t = 0$ is even stronger for aggregate time-series data encompassing a sufficiently long period.[3]

Equation (2) can be used to estimate the value of r. Inasmuch as economic conditions in the United States have varied over the sample period of 1929–69, r cannot be expected to have remained constant over the years; after all, it is a form of weighted average of observable and nonobservable interest rates. The observable rates have fluctuated widely with cyclical fluctuations in economic activity, and there may be little reason to assume that the nonobservable rates, that are mainly dependent on the degree of optimism and pessimism of the household units, have remained constant.

Since some sequential variations in r may be of the transitory type and others of a more permanent nature, we use the varying parameters regression procedure to obtain annual estimates of r. This method which has recently been developed by Cooley and Prescott (1973*a*, 1973*b*) has two advantages. On the one hand this approach is particularly suitable in correcting for autocorrelation and on the other it incorporates the effects of structural changes into the consumer discount rate.

---

[3] Treating the mean value of $\bar{y}_t = 0$ has become a common practice (see, for example, Goldberger 1964; Walters 1970; Johnston 1972, and Darby 1974).

TABLE 1

TIME SERIES ESTIMATES OF *r*, THE SUBJECTIVE INTEREST RATES

Year	*r*	Year	*r*	Year	*r*
1929	.25	1943	.28	1957	.25
1930	.23	1944	.28	1958	.25
1931	.22	1945	.27	1959	.26
1932	.19	1946	.24	1960	.26
1933	.19	1947	.24	1961	.26
1934	.20	1948	.24	1962	.26
1935	.21	1949	.24	1963	.26
1936	.24	1950	.24	1964	.27
1937	.24	1951	.25	1965	.27
1938	.23	1952	.25	1966	.28
1939	.24	1953	.25	1967	.28
1940	.25	1954	.25	1968	.28
1941	.26	1955	.26	1969	.28
1942	.28	1956	.26		

Table 1 presents the time-series estimates of *r*. They require some explanations. First, the estimates, as expected, are not constant over the sample period. Second, they are extremely high relative to the market interest-rate structure. Friedman (1963, p. 25) points out, "If it [the subjective interest rate] seems drastically out of line with the widely quoted market rates of interest, it should be kept in mind that these rates apply only to a very limited range of assets; that most future receipts whose discounted value constitutes wealth come from assets that cannot be readily bought and sold or for which buying and selling price differ widely. It should be kept in mind also that the value of *r* required by the theory [PIH] includes not only pecuniary but also non-pecuniary returns from the assets. These non-pecuniary returns rationalize the simultaneous ownership by consumer unit of assets that have widely varying rates of pecuniary yields."

Third, the consumer horizon is very sensitive to socioeconomic conditions. For instance, in 1929 where $r = .25$ the consumer horizon was 4 years in length. However, as depression ensued *r* declined rapidly to .19 in 1932 and 1933, and consumers became more cautious by lengthening their horizon to 5.26 years, a 32 percent increase. As depression receded the length of the horizon fell back to 4 years in 1940. In the years following World War II and through the decade of the fifties the consumer horizon remained on the average around 4 years. During the sixties, especially toward the latter part of the decade, the consumer horizon shrank to 3.57 years, showing a more confident attitude on the part of the consumer.

Although our results do not show an average consumer horizon of exactly 3 years, they are sufficiently close to lend support to Friedman's position.

Estimation of the Factor of Proportionality, k

Friedman defines k as the MPC = APC that relates permanent consumption to permanent income as

$$C_p = kY_p \tag{3}$$

where $Y_p = \hat{r}W$ is permanent income, $\hat{r}$ is the estimated value of r, and C_p is permanent consumption. Since C_p is not observable we follow Friedman (1957, p. 22) in substituting its equivalent as

$$C_m - C_t = kY_p \quad \text{or} \quad C_m = kY_p + C_t, \tag{4}$$

with $\bar{C}_t = 0$ and C_m, C_t, and $\bar{C}_t$ as measured consumption, transitory consumption, and the mean value of transitory consumption, respectively. Equation (4) states that actual consumption differs from permanent consumption by the amount of transitory consumption. This difference is mainly due to "accidental" or "chance" factors coupled with errors of measurement. It is expected that the average value of the transitory consumption will be zero over the sample period (Friedman 1957, p. 22). In estimating k we have abandoned the assumption of parametric stability and have allowed k to be subject to permanent and transitory changes over time. Using the varying parameters regression technique the time-series magnitudes of k are presented in figure 1, which reveals some interesting points.

To begin with, the overall estimated value of $k = MPC = APC$ is equal to .72. This figure seems to be somewhat lower than the marginal propensity to consume that has been obtained by Friedman and others. However, there are two reasons for the discrepancy of the results. (1) Most studies, due to lack of data on total wealth, have used weighted averages of the past and present disposable income to calculate permanent income whereas we have derived permanent income directly from wealth. Therefore our income concepts are not comparable. (2) The definition of consumption as employed here refers to the consumer expenditures on nondurable goods plus the services of durable goods. Other studies, due to the unavailability of data, have used either total consumer expenditures or consumer expenditures on nondurables only.

Although the overall behavior of k is stable when economic conditions are normal its value is not constant from year to year.[4] In the postwar period, where economic conditions were relatively stable, the fluctuations in k were minor. However, when looking at more turbulent periods the value of k has varied substantially; for instance, during the depres-

---

[4] The temporal variation in k is in line with Friedman's intuitive view, when he referred to a number of studies, that "the observed rough constancy of k is about as much of a puzzle as substantial variations in it would be" (1957, p. 120).

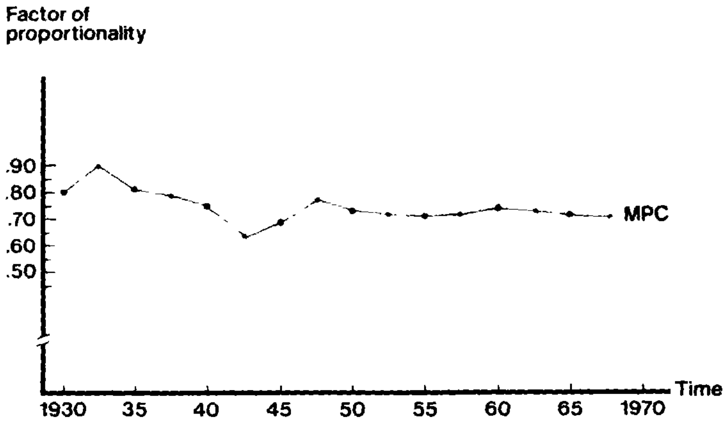


FIG. 1

sion years it rose from .75 in 1929 to .89 in 1933 —a 19 percent increase. With the economic recovery the value gradually tapered off, reaching .75 in 1940. The war years produced a sizable decline to a low of .63—a 16 percent decline from 1940.

To obtain a clear view of the fluctuation of k in an unsettled economic environment we search for clues in the behavior of its determinants. The theory states that

$$k = k(i, w, u), \tag{5}$$

with $\partial k/\partial i < 0$, $\partial k/\partial w > 0$, and $\partial k/\partial u \lesseqgtr 0$ where i is the market interest rate structure, w is the ratio of nonhuman wealth to human wealth or to income,, and u is a catchall phrase representing numerous traits and attributes of the household and such factors that affect their degree of optimism.

During the Great Depression the nominal as well as the real interest rate structure, whether represented by the discount rate or the prime rate, declined between 50 and 70 percent. The effects of i on k according to the theory is to raise the value of k. By the same token, the value of w, the ratio of nonhuman wealth to income, rose during this period. Although during the depression both income and nonhuman wealth declined, the reduction in income, as evidenced by a high unemployment rate which reached a peak of 25 percent in 1933, was more severe. On the other hand, total nonhuman wealth, the numerator of w, which consists of tangible assets and net financial claims against governments and the rest of the world, declined less rapidly than income. The increase in w, according to the theory, exerted an upward influence on k.

The effects of u upon k are not clearly known. However, during the depression a number of welfare programs had been instituted and the federal government pursued an expansionary policy culminating in a deficit spending of $2.5 billion in 1932. By the same token, as depression reached the second leg of its double bottom in March 1933 signs were evident that panic was terminated and psychological reversal had taken place (Lee 1971, p. 189). This was due partly to the experimental policies of the Roosevelt administration which were aimed at the economic recovery. The establishment of the National Recovery Administration is a good example of the administration effort to build public confidence. It may be reasonable to expect that the effects of u on k were positive. In sum, the overall effect of i, w, and u on k was to raise the value of k during the depression years.

On the other hand, the value of k declined sharply during World War II. This may be explained mainly in terms of the federal government's programs to finance the war, diverting resources from the private sector to the public sector.[5]

Verifications of PIH Postulates

Finally, we attempt to verify the remaining postulates of the PIH. The theory states that the transitory components of income and consumption are uncorrelated with one another and with the corresponding permanent components. Using the correlation coefficient, R, this relation is expressed symbolically as

$$R(C_t, y_t) = R(C_p, C_t) = R(y_p, y_t) = 0. \qquad (6)$$

Our estimates of the coefficient of determination, R^2, for each pair are as follows: $R^2(C_t, y_t) = .11$, $R^2(C_p, C_t) = .08$, $R^2(y_p, y_t) = .02$. These results strongly favor the PIH that there is no correlation between the transitory elements and the transitory and permanent elements.

The theory also states that the only systematic relationship that exists is between permanent consumption and permanent income. Our findings of $R^2 = .99$ between C_p and y_p support this hypothesis.

II. Conclusion

In this study we have provided new evidence on consumer subjective discount rate and the consumer horizon. Our findings support Friedman's 3-year average horizon. We have also estimated a time series of the factor of proportionality and have analyzed its behavior over the sample period. Finally, we have verified the correlation between the permanent elements, permanent and transitory elements, and the transitory elements. The results overwhelmingly support Friedman.

---

[5] During the war the government pursued an aggressive campaign that was directed at individual savers to buy government bonds. Though we do not know the exact figures of diversion of resources from the private sector to the government sector for the war years it was well above the 18 percent which is the yearly average resource diversion for the entire period 1939–49 (Friedman 1952, pp. 624–25).

References

Christensen, L. R., and Jorgenson, D. W. "U.S. Income, Saving and Wealth, 1929–1969." *Rev. Income and Wealth* 19 (December 1972): 329–62.

Cooley, T. F., and Prescott, E. C. "An Adaptive Regression Model." *Internat. Econ. Rev.* 14 (June 1973): 364–71. (*a*)

———. "Systematic (Non-Random) Variation Models Varying Parameter Regression: A Theory and Some Applications." *Ann. Econ. and Soc. Measurement* 2 (January 1973): 463–73. (*b*)

Darby, Michael. "The Permanent Income Theory of Consumption—a Restatement." *Q.J.E.* 88 (May 1974): 228–50.

Friedman, Milton. "Price, Income, and Monetary Changes in Three Wartime Periods." *A.E.R.* 42 (May 1952): 612–25.

———. *A Theory of the Consumption Function.* Princeton, N.J.: Princeton Univ. Press, 1957.

———. "Windfalls, the 'Horizon' and Related Concepts in the Permanent Income Hypothesis." In *Measurement in Economics: Studies in Mathematical Economics and Econometrics,* edited by Carl F. Christ et al. Stanford, Calif.: Stanford Univ. Press, 1963.

Goldberger, A. S. *Econometric Theory.* New York: Wiley, 1964.

Griliches, Z.; Maddala, G.; Lucas, R.; and Wallace, N. "Notes on Estimated Aggregate Quarterly Consumption Function." *Econometrica* 30 (July 1962): 491–500.

Holbrook, R. S. "The Three Year Horizon: An Analysis of the Evidence." *J.P.E.* 75, no. 5 (October 1967): 750–54.

Johnston, J. *Econometric Methods.* New York: McGraw-Hill, 1972.

Landsberger, M. "Consumer Discount Rate and the Horizon: New Evidence." *J.P.E.* 79, no. 6 (November/December 1971): 1346–59.

Laumas, P. S. "A Test of the Permanent Income Hypothesis." *J.P.E.* 77, no. 5 (September/October 1969): 857–61.

Laumas, P. S., and Mohabbat, K. A. "The Permanent Income Hypothesis: Evidence from Time-Series Data." *A.E.R.* 62 (September 1972): 730–34.

Lee, M. W. *Macroeconomics: Fluctuations, Growth, and Stability.* Homewood, Ill.: Irwin, 1971.

Walters, A. A. *An Introduction to Econometrics.* New York: Norton, 1970.

Wright, C. "Estimating Permanent Income: A Note." *J.P.E.* 77, no. 5 (September/October 1969): 845–50.

Equilibrium in Stable Markets

James A. Ohlson

University of California, Berkeley

E. Fama (1971) has shown that the classical, two-period, two-parameter capital asset pricing model can be generalized to the case of symmetric stable distributions. Fama develops his results using a one-factor ("market-model") distribution of returns. The present research shows that the same equilibrium results hold for all stable distributions of returns which allow for a concave and differentiable objective function; there is no need to assume symmetry or any other restrictions on the return structure. Furthermore, derivations are straightforward in that they rely only on elementary properties of homogeneous functions.

I

Fama (1971) has developed a model which generalizes the capital asset pricing model of Sharpe (1964), Lintner (1965), and Mossin (1966). Specifically, Fama analyzes symmetric stable (Paretian) return distributions with the characteristic exponent (α) greater than 1, producing equilibrium structures which are similar in most respects to the classical capital asset pricing model.

The analysis set forth by Fama employs a one-factor market model which causes the interdependence of asset returns. One may, of course, hypothesize that the results can be generalized and do not depend on this specification of the return process. Fama correctly suggests an extension based on specifying a vector of commonality factors. Therefore, one would expect the demand and equilibrium analysis to remain essentially unchanged for *all* multivariate symmetric stable distributions with $\alpha > 1$ without complicating the formal analysis.

It will be shown below that this is true. The analysis can be made simple, efficient, and clear by using elementary properties of homogeneous

[*Journal of Political Economy*, 1977, vol. 85, no. 4]
© 1977 by The University of Chicago. All rights reserved.

functions. If the derivations are cast in this mode, it will become apparent that the results are not contingent upon the assumption that returns have a symmetric distribution. They are equally applicable for asymmetric stable distributions, as long as the portfolio-optimization problem is concave in the allocation vector. The equilibrium relationships derived by Fama are thus of considerable generality; consequently, it is interesting to build a framework which allows for efficient and general derivations.

II

The set of all possible multivariate symmetric stable distributions admits a general representation by their characteristic functions. Let $\tilde{\mathbf{R}} \equiv (\tilde{R}_1, \ldots, \tilde{R}_n)'$ be an n-dimensional vector of random variables obeying such a probability law. Ferguson (1955) has shown that $\tilde{\mathbf{R}}$ is symmetric multivariate stable if and only if its characteristic function is representable in a form given by[1]

$$\log E \exp \{i\mathbf{t}'\tilde{\mathbf{R}}\} = i\delta(\mathbf{t}) - \sigma(\mathbf{t})^{\alpha}, \tag{1}$$

where $\mathbf{t} \in E^n$, $i = \sqrt{-1}$, $0 < \alpha \leq 2$ and for all real scalars u

$$\delta(\mathbf{t}u) = u\delta(\mathbf{t}), \tag{1a}$$

$$\sigma(u\mathbf{t})^{\alpha} = |u|^{\alpha}\sigma(\mathbf{t})^{\alpha} \geq 0. \tag{1b}$$

Hence the conditions (1a) and (1b) assert that $\delta(\mathbf{t})$ and $\sigma(\mathbf{t})$ are functions which are homogeneous of degree 1. This is an important fact which will be extensively used in the following analysis. Ferguson further demonstrated (1955, p. 149) that (1a) can be reduced into a solution $\mathbf{t}'\boldsymbol{\delta}$, where $\boldsymbol{\delta} \equiv (\delta_1, \ldots, \delta_n)' \in E^n$, and $\boldsymbol{\delta}$ is a vector of scalars independent of $\mathbf{t}$. No similar simplification is available for $\sigma(\mathbf{t})$; for example, if $\alpha = 2$ (i.e., $\tilde{\mathbf{R}}$ has a normal distribution) $\sigma(\mathbf{t}) = [\sum \sum \frac{1}{2} t_i t_j \operatorname{cov}(\tilde{R}_i, \tilde{R}_j)]^{1/2}$.

III

Below, we shall consider only the case in which $1 < \alpha \leq 2$ and $\tilde{\mathbf{R}}$ should be interpreted as the vector of 1-period rates of return on the n-risky assets in the economy. Next, we can consider the type of economy described in Fama (1971) without imposing any particular restriction on the symmetric, multivariate, return distribution. In other words, the economy is a standard "2-period" frictionless model where the investors or consumers have identical probability beliefs and every investor maximizes his expected utility of the 2-period consumption streams, c_1 and $\tilde{c}_2$. Private borrowing and lending at the same rate is permitted, and the rate is denoted by R_f.

---

[1] See Ferguson, p. 147. It may also be noted that Ferguson's parameterization is somewhat different from the one given here.

Using a similar notation as in Fama (1971), optimal behavior for an investor can be represented by max $EU(c_1, \tilde{c}_2) \equiv$ max $V[c_1, \boldsymbol{\delta}'\mathbf{x} + R_f x_{n+1}, \sigma(\mathbf{x})]$, where

$$\tilde{c}_2 = (w_1 - c_1)(1 + \sum^{n} x_i \tilde{R}_i + R_f x_{n+1}), \qquad \mathbf{x} \in E^n,$$

and

$$\sum^{n+1} x_j = 1;$$

the decision variables $x_1, \ldots, x_{n+1}$ represent the relative portfolio weights. The wealth not consumed in the first period, $w_1 - c_1$, is invested and the rate of return is given by a two-parameter distribution with mean $\mathbf{x}'\boldsymbol{\delta} + x_{n+1}R_f$ and dispersion parameter $\sigma(\mathbf{x})$. It is further assumed that V(or U) satisfies mild regularity conditions to the extent that it is concave and differentiable in the relevant arguments.

Differentiating V with respect to the weights $x_1, \ldots, x_{n+1}$ yields the optimality conditions

$$V_2\delta_i + V_3\sigma_i(\mathbf{x}) = \lambda, \qquad i = 1, \ldots, n$$

$$V_2 R_f = \lambda,$$

where $V_2 \equiv \partial V(\cdot, y, \cdot)/\partial y$, $V_3 \equiv \partial V(\cdot, \cdot, y)/\partial y$, and $\sigma_i(\mathbf{x}) \equiv \partial\sigma(\mathbf{x})/\partial x_i$, and λ is the Lagrangian multiplier. A further reduction of the above gives

$$\frac{\delta_i - R_f}{\delta_j - R_f} = \frac{\sigma_i(\mathbf{x})}{\sigma_j(\mathbf{x})}, \qquad i = 1, \ldots, n. \tag{2}$$

The function $\sigma(\mathbf{x})$ is homogeneous of degree 1, and this entails that $\sigma_i(\mathbf{x})$ is homogeneous of degree 0;[2] that is, $\sigma_i(u\mathbf{x}) = \sigma_i(\mathbf{x})$. Consequently, the optimal allocation vector is generated by $k(V, w_1)\mathbf{x}^*$, where $\mathbf{x}^*$ is the vector that satisfies (2) subject to

$$\sum_{j=1}^{n} x_j^* = 1.$$

---

[2] The proof is as follows. Let $\Delta\mathbf{x} \equiv (0, \ldots, 0, \Delta x_i, 0, \ldots, 0)'$ and $\sigma_i(\mathbf{x}) \equiv \partial\sigma(\mathbf{x})/\partial x_i$; then

$$\begin{aligned}
\sigma_i(k\mathbf{x}) &= \lim_{\Delta x_i \to 0} [\sigma(k\mathbf{x} + \Delta\mathbf{x}) - \sigma(k\mathbf{x})]/\Delta x_i \\
&= \lim_{\Delta x_i \to 0} k[\sigma(\mathbf{x} + k^{-1}\Delta\mathbf{x}) - \sigma(x)]/k(k^{-1}\Delta x_i) \\
&= \lim_{k^{-1}\Delta x_i \to 0} [\sigma(\mathbf{x} + k^{-1}\Delta\mathbf{x}) - \sigma(\mathbf{x})]/(k^{-1}\Delta x_i) \\
&= \sigma_i(\mathbf{x}).
\end{aligned}$$

The usual type of portfolio-separation result can be seen producing identical optimal mixes in the risky assets for all investors in the economy.[3] The V and w_1 represent circumstances unique to the investor and can be dismissed.

The equilibrium relationship relating risk to expected return is now easily derived. Using Euler's theorem, one has

$$\sigma(\mathbf{x}) = \sum^{n} x_i \sigma_i(\mathbf{x}) \qquad \text{for all } \mathbf{x} \in E^n,$$

since $\sigma(\mathbf{x})$ is homogeneous of degree 1. In particular, putting $\mathbf{x} = \mathbf{x}^*$ and applying Euler's theorem to expression (2), one gets

$$\frac{1}{\delta_j - R_f} \sum_{i=1}^{n} x_i^*(\delta_i - R_f) = \frac{1}{\sigma_j(\mathbf{x}^*)} \sum_{i=1}^{n} x_i^* \sigma_i(x^*) = \frac{\sigma(\mathbf{x}^*)}{\sigma_j(\mathbf{x}^*)}.$$

Write

$$\sum_{i=1}^{n} x_i^* \delta_i \equiv \delta_M$$

and the last expression simplifies to

$$\delta_j - R_f = \sigma_j(\mathbf{x}^*)\left[\frac{\delta_M - R_f}{\sigma(\mathbf{x}^*)}\right]. \tag{3}$$

Expression (3) corresponds to Fama's equation 31 (1971, p. 46); the derivations here are simpler and more general; otherwise no difference is present. Equation (3) can also be written as

$$\delta_j - R_f = \frac{1}{\alpha} \frac{\partial \sigma(\mathbf{x}^*)^\alpha}{\partial x_j} \left[\frac{\delta_M - R_f}{\sigma(\mathbf{x}^*)^\alpha}\right], \tag{4}$$

and for $\alpha = 2$ this is clearly the Sharpe-Lintner-Mossin capital asset pricing model.

The above analysis has been stated in terms of returns rather than a vector of time 1 asset prices and a vector of time 2 random gross payoffs. To change the analysis to the latter type of setting is not difficult. It leads one then to the valuation equations 34 and 35 in Fama's paper.

IV

Certain final comments are now appropriate. First, it may seem that formal derivations should be applicable even if $\alpha \leq 1$. This is not quite true, however. Now it is not possible to formulate a well-behaved expected utility problem: $\alpha \leq 1$ implies $E \sum x_j \tilde{R}_j$ is not finite and, for such symmetric distributions, one cannot use a utility function which is concave in second-period consumption ($\tilde{c}_2$). Expected utility is no longer finite if $x_j \neq 0$ for at least one $j (j = 1, \ldots, n)$.

---

[3] This is the type of separation originally developed by Tobin (1958); he derived the results for the case in which returns are normal.

Second, it should also be stressed that multivariate symmetric distributions need not be represented or constructed by the market-factor(s) models which Fama uses in his formal analysis.[4] For example, it may be shown that if $\sigma(t)$ in (1) equals $\sigma(\mathbf{t})^\alpha = [(\mathbf{t}'A_1\mathbf{t})^{\alpha_1/2} + (\mathbf{t}'A_2\mathbf{t})^{\alpha_1/2}]^{\alpha_2}$, then this represents a multivariate symmetric stable distribution if $0 < \alpha_2 \leq 1$ and $0 < \alpha_1 \leq 2$; A_1 and A_2 are any two positive semidefinite symmetric matrices, and the characteristic exponent is equal to $\alpha_1\alpha_2$.[5]

Finally, one might conjecture that (3) should hold as a general equilibrium condition for *all* two-parameter multivariate stable distributions which can be combined with concave utilities. That is, the assumption of symmetry has no particular significance in the analysis. This would be true in all cases but for the special one of $\alpha = 1$. Nevertheless, the formal analysis will change little. In the above-mentioned Ferguson paper, he demonstrates how (1) is modified to represent all multivariate stable distributions. Most important, it is still true that portfolio returns (or positive linear transformations of $\tilde{\mathbf{R}}$) have (*a*) a location parameter given by $\sum \delta_i x_i$, and (*b*) a dispersion parameter with the same homogeneity property as described by expression (1b). Therefore, all of the formal two-parameter analyses will remain unchanged. When a well-behaved maximization problem can be specified, the optimality condition continues to be given by (2), and the second-order conditions are automatically satisfied because it is assumed that V is concave in the portfolio allocation vector. Quite simply, the reader can verify directly that all of the Fama analysis will carry through even if it is assumed that the components of the market-factor model are comprised of stable variates with identical nonzero skewness parameters. Naturally, any geometrical analysis may change slightly when one can no longer presume that the dispersion parameter is a "bad good" (which is the case whenever the returns are symmetric). In conclusion, the two-parameter equilibrium model in stable markets derived by Fama has great generality. Since the analysis depends entirely on properties of homogeneous functions, assumptions can be relaxed completely without introducing any mathematical complications.

References

Fama, Eugene F. "Risk, Return, and Equilibrium." *J.P.E.* 79, no. 1 (January/February 1971): 30–55.

---

[4] Press (1972) discusses multivariate stable distributions.

[5] A formal proof is straightforward but fairly lengthy. It will be supplied by the author if requested.

Ferguson, T. *On the Existence of Linear Regression in Linear Structural Relations.* University of California Publications in Statistics, vol. 2, no. 7. Berkeley: Univ. California Press, 1955.

Lintner, J. "The Valuation of Risk Assets and the Selection of Risky Investments in Stock Portfolios and Capital Budgets." *Rev. Econ. and Statis.* 47 (February 1965): 13–37.

Mossin, J. "Equilibrium in a Capital Asset Market." *Econometrica* 34, no. 4 (October 1966): 768–83.

Press, S. J. "Multivariate Stable Distributions." *J. Multivariate Analysis* 2 (1972): 444–62.

Sharpe, W. F. "Capital Asset Prices: A Theory of Market Equilibrium under Conditions of Risk." *J. Finance* 19, no. 4 (September 1964): 425–42.

Tobin, J. "Liquidity Preference as Behavior towards Risk." *Rev. Econ. Studies* 25 (February 1958): 65–86.

On the Shadow Pricing of Traded Commodities

Peter G. Warr

Monash University, Melbourne

This paper extends the case for shadow pricing traded commodities at their relative international prices in benefit-cost analysis. This result is shown to hold (*a*) when there are nontraded commodities whose (possibly distorted) prices are indirectly affected by public production of traded commodities, and (*b*) when there is a government budgetary constraint. Contrary to arguments found in the literature, neither of these cases in itself provides an argument for shadow pricing traded commodities at values other than their relative international prices.

Several studies have recently attempted to provide guidelines and general principles for the calculation of shadow prices for benefit-cost analysis, the most important being Little and Mirrlees (1969, 1974) and Dasgupta, Marglin, and Sen (1972). These studies typically recommend that benefit-cost analysts value traded commodities relative to one another at their respective international prices, even though there are distortionary tariffs and other taxes present in the domestic markets for these and other commodities. Since these studies are addressed primarily to practitioners of benefit-cost analysis, however, they are understandably deficient in their specification of the economic model underlying the advocated shadow pricing rules. This has led to confusion about the analytical basis of shadow pricing and criticism of the shadow pricing procedures recommended.

It is well known that the "first-best" optimal pricing rule for traded commodities is that *all* producers in the economy set their marginal rates of transformation between these commodities at their relative inter-

This paper has benefited from the comments of Partha Dasgupta, George Feiger, and an anonymous referee. The research was supported by a grant from USAID to the Economic Development Center of the University of Minnesota. The author is responsible for all views and any errors it contains.

[*Journal of Political Economy*, 1977, vol. 85, no. 4]

© 1977 by The University of Chicago. All rights reserved.

national prices. But the existence of fixed, distortionary tariffs and other taxes on these commodities means that some of the necessary conditions for a first-best optimum will not be met, namely, those relating to private firms. If project evaluators must take the existence of these taxes as given, then the distortions concerned must be considered as constraints for the shadow pricing exercise they perform. Shadow pricing must therefore be viewed as an exercise in the economics of the second-best. The question now is: Are international prices indeed the correct second-best shadow prices for the relative valuation of traded commodities? This has been denied by several authors, including Rudra (1972) and Weckstein (1972).

Weckstein criticizes the valuation of traded commodities at their relative international prices on the grounds that these are "optimal prices" and are therefore "irrelevant to an economy that does not in fact achieve an optimal allocation of resources" (p. 479). He argues that "when a price rule is to be applied partially, the objective should be to find prices that signal an optimal accommodation to a given non-optimal structure of prices and allocations—a 'second-best' criterion" (p. 478). The correct second-best shadow prices, Weckstein implies, would take account of the indirect effects of public production of traded commodities on the market prices of nontraded commodities, although it is the case that if there were no nontraded commodities the shadow pricing of traded commodities at their relative international prices would be appropriate. Weckstein recommends that, in practice, traded commodities be valued at their relative domestic market prices, even though these are known to be distorted by tariffs.

An analysis based on a more fully specified model was recently presented by Dasgupta and Stiglitz (1974), who recommend the following shadow pricing rules: R_1, "The shadow price of a tradeable commodity is its international price unless . . . there is a government budgetary constraint . . ." and R_2, "When there is a government budgetary constraint, the shadow price of a tradeable lies between the world price and its domestic price" (p. 30). They claim that these results hold even when there are fixed, nonoptimally chosen tariffs and other taxes (p. 30), but unfortunately the Dasgupta–Stiglitz analysis of this case assumes that all commodities are traded.[1] The applicability of the result is thus limited by this assumption. A subsequent paper by Boadway (1975) analyzes this problem in a model which contains nontraded commodities, but the analysis is limited first by the failure to consider distortions in the private sector explicitly as constraints on the public sector's shadow pricing exercise, and second by the assumption that public production of traded commodities has no indirect effects on the prices of nontraded commodities. Consequently, neither the Dasgupta–Stiglitz nor the Boadway analysis adequately answers Weckstein's objections.

---

[1] This analysis is found in section 5.1 of Dasgupta and Stiglitz (1974). Rules R_1 and R_2 are rules 2 and 3, respectively, of Dasgupta and Stiglitz.

Section I of this paper presents a simple model for the analysis of the above issues. In Section II we show that the existence of nontraded commodities whose (possibly distorted) prices may indirectly be affected by the public production of traded commodities does not change the result that traded commodities should be valued at their relative international prices. This provides an important extension of rule R_1 of Dasgupta and Stiglitz. Section III shows that rule R_2 of Dasgupta and Stiglitz is misleading and based on special unstated assumptions. We show that, except in a special case, traded commodities should be valued at their relative international prices irrespective of whether there is a government budgetary constraint.[2]

I. The Model

We consider an economy in which there is a single consumer and two producers, one "public" and the other "private." The private firm maximizes profits at domestic market prices, and the public project maximizes shadow-profit using the shadow prices given it by a planner. Our concern is with the optimal determination of these shadow prices. The consumer possesses a quasi-concave utility function $U(c)$ where $c = (c_1, c_2, \ldots, c_k)$ is his net consumption vector. The private firm has the concave implicit production function $F(y) = 0$ where $y = (y_1, y_2, \ldots, y_k)$ is the firm's net production vector. The public project's concave implicit production function is written $G(x) = 0$ where $x = (x_1, \ldots, x_k)$ is the project's net production vector. Note that the consumer's utility function and both production functions are defined on the same set of commodities: T is the set of traded commodities, and N is the set of nontraded commodities.

The trade balance constraint for this economy is written

$$\sum_{t \in T} r_t(x_t + y_t - c_t) \geq 0 \tag{1}$$

where r_t is the international price for traded commodity t, and the commodity balance constraint is written

$$x_n + y_n - c_n \geq 0, \qquad \text{all } n \in N. \tag{2}$$

---

[2] It is important to realize that the Dasgupta–Stiglitz propositions above, and the results of this paper, relate to the *relative* shadow prices of traded commodities, since a traded commodity is chosen as numeraire. Consequently, the reader should not interpret Dasgupta–Stiglitz, Boadway, or the present paper to have shown that the shadow price of foreign exchange is necessarily equal to the prevailing market rate. In the presence of distortionary taxes this will typically not be the case.

The first-best optimization problem is now max $U(c)$ subject to (1), (2), $F(y) = 0$, and $G(x) = 0$. We formulate the Lagrangian

$$\begin{aligned} L^* = U(c) + \lambda\Big[\sum_{t \in T} r_t(x_t + y_t - c_t)\Big] \\ + \sum_{n \in N} \rho_n(x_n + y_n - c_n) + \mu G(x) + \gamma F(y) \end{aligned} \tag{3}$$

where λ, ρ_n, μ, and γ are Lagrangian multipliers. The necessary conditions for a first-best optimum are well known. Using commodity 1 (a traded commodity) as numeraire and confining ourselves to interior solutions, we have, for traded commodities t and 1, $G_t^*/G_1^* = F_t^*/F_1^* = U_t^*/U_1^* = r_t/r_1$, t, $1 \in T$, and for the nontraded commodity n, $G_n^*/G_1^* = F_n^*/F_1^* = U_n^*/U_1^* = \rho_n^*/\lambda^* r_1$, $n \in N$, where * denotes evaluation of the quantity concerned at this first-best optimum.

II. "Second-best" Shadow Prices for Traded Commodities

We now introduce distortions in domestic market prices as follows. Let the ruling domestic market prices be denoted $p_1^0, p_2^0, \ldots, p_k^0$. We assume that any distortions in these prices are due to fixed, nonoptimal tariffs and other taxes and that, from the point of view of the planner concerned, these must be taken as given and fixed.[3] Then profit maximization of the private firm gives, in the case of traded commodities,

$$\frac{F_t}{F_1} = \frac{p_t^0}{p_1^0}, \tag{4}$$

and in the case of nontraded commodities[4]

$$\frac{F_n}{F_1} = \frac{p_n^0}{p_1^0}. \tag{5}$$

Obviously, $p_t^0/p_1^0 = r_t/r_1$ and $p_n^0/p_1^0 = U_n^*/U_1^*$ when there are no distortionary tariffs or other taxes in the markets concerned and differ from these values otherwise. We now reformulate the basic Lagrangian equation (3) to incorporate these distortions as constraints:

$$\begin{aligned} L^{**} = U(c) + \lambda\Big[\sum_{t \in T} r_t(x_t + y_t - c_t)\Big] \\ + \sum_{n \in N} \rho_n(x_n + y_n - c_n) + \mu G(x) + \gamma F(y) \\ + \sum_{i \in T \cup N} \alpha_i(p_i^0/p_1^0 - F_i/F_1), \end{aligned} \tag{6}$$

---

[3] The first of these assumptions is made throughout Dasgupta and Stiglitz (1974), while the second corresponds to their section 5.

[4] Note that the private firm behaves as a perfect competitor, treating market prices parametrically. Note also that we confine ourselves to interior solutions, as we do throughout this paper.

where α_i is a Lagrangian multiplier and $\alpha_i = 0$ when p_i^0/p_1^0 does not differ from its optimal value.

We now differentiate (6) with respect to x_t and x_1 as before to obtain the rate of transformation between traded commodities in the public project at the second-best optimum (indicated by $**$),

$$\frac{G_t^{**}}{G_1^{**}} = \frac{\lambda^{**} r_t + \sum_i \alpha_i^{**} \partial(p_i^0/p_1^0)/\partial x_t^{**}}{\lambda^{**} r_1 + \sum_i \alpha_i^{**} \partial(p_i^0/p_1^0)/\partial x_1^{**}}. \tag{7}$$

The question now is: Does equation (7) reduce, under our assumptions, to $G_t^{**}/G_1^{**} = r_t/r_1$?

Consider the summation terms appearing in the numerator and denominator. These expressions involve the effects of public production on relative producer prices. Since we assume that international prices are given, and that all distortions are due to nonoptimal tariffs and other taxes, the relative producer prices of traded commodities can be affected by public production only if tariff levels are affected. These effects are ruled out by the assumption that taxes are fixed,[5] so for all $i \in T$ the terms in the summation expressions concerned are zero. Now consider the terms $\partial(p_w^0/p_1^0)/\partial x_t^{**}$ (where $w \in N$; t, $1 \in T$). We note that in this model public production (use) of traded commodities can affect the producer prices of nontraded commodities relative to traded commodities in only two ways: first, it could affect the levels of taxes on those commodities; second, it could affect these relative prices via the general equilibrium effects of its contributions to (depletion of) foreign exchange earnings. The first effect is ruled out, so we now consider the second.

If we define f_x to be the net contribution of public production to total foreign exchange earnings, then

$$f_x = \sum_{t \in T} r_t x_t, \tag{8}$$

and it follows that

$$\frac{\partial(p_w^0/p_1^0)}{\partial x_t^{**}} = \frac{\partial(p_w^0/p_1^0)}{\partial f_x^{**}} \frac{\partial f_x^{**}}{\partial x_t^{**}} = \frac{\partial(p_w^0/p_1^0)}{\partial f_x^{**}} r_t. \tag{9}$$

This implies that

$$\frac{\partial(p_w^0/p_1^0)/\partial x_t^{**}}{\partial(p_w^0/p_1^0)/\partial x_1^{**}} = \frac{r_t}{r_1}, \tag{10}$$

and equation (7) reduces immediately to $G_t^{**}/G_1^{**} = r_t/r_1$. Thus, when all market distortions are due to nonoptimal taxes, and the levels of these taxes are unaffected by public production of traded commodities, relative international prices are still the correct shadow prices for traded commodities in the public sector. This result holds despite the fact that private sector producers face nonoptimal (tax-distorted) prices both for these commodities and for nontraded commodities and despite the fact that the prices of nontraded commodities may indirectly be affected by public production of traded commodities.

---

[5] See section 6 of Dasgupta and Stiglitz (1974) and Sen (1972) for discussions of this issue.

III. The Implications of a Government Budgetary Constraint

Suppose there is a government budgetary constraint, meaning, as Dasgupta and Stiglitz put it, "a constraint on the government budget deficit in any given year" (p. 28). We now consider its implications for public projects. When a public project buys or sells a commodity it affects the government's budgetary deficit in two ways: (1) through the financial balance sheet of public projects, and (2) through the generation of tax revenues. For example, suppose a project buys a tractor (a traded commodity) on the domestic market from an import agent. The effect of this on the government's budget deficit is the domestic price paid *minus* the effect of the transaction on tax revenues—the tariff and any sales tax. The latter items constitute both an expenditure and a receipt for the public sector as a whole; at most these taxes are simply an intragovernmental transfer of funds, and their payment by projects has no effect on the size of the government's budgetary deficit. In formulating the budgetary constraint actually facing public sector production we can make either of two assumptions: A_1, the effects of project activity on the generation of tax revenues are recognized; or A_2, these effects are ignored. Under A_1 the budget constraint would be formulated $\hat{p}^0 \cdot x \geq b$, where $\hat{p}^0$ is the vector of domestic prices *net* of the effects that public use (production) of the various commodities has on the generation of tax revenues. In the case of nontraded commodities this involves the general equilibrium effects of public sales or purchases of those commodities, but in the case of traded commodities the matter is more straightforward. Since it has been assumed that the only distortions in the economy are those due to tax policy it must be true that, for traded commodities t and 1, $\hat{p}^0_t/\hat{p}^0_1 = r_t/r_1$, the ratio of their international prices.

Case A_2 implies that the budgetary constraint would be formulated $p^0 \cdot x \geq b$, where p^0 is simply the vector of domestic prices actually paid and received by public projects.[6] Since this is the formulation that Dasgupta and Stiglitz employ, it seems that A_2 is what they have in mind, although this assumption is not made explicit by them. In practical circumstances, this myopic kind of budgetary constraint may frequently be the kind actually faced by public projects, but this is not really "a constraint on the government budget deficit," as Dasgupta and Stiglitz state. It is a constraint on the financial deficit of public projects expressed in domestic market prices, and this constraint may exist with or without the presence of a government budgetary constraint. The two are quite distinct.

---

[6] We assume, of course, that public projects actually pay the taxes and tariffs levied on the commodities concerned. Otherwise the results of this section follow trivially.

Now consider the implications of A_1. We denote the dual variable associated with the constraint corresponding to A_1 by δ, append this new constraint to our basic Lagrangian (eq. [6]) and derive the first-order conditions for optimal public production of the tradeables t and 1 at this new "third-best" optimum. Again confining ourselves to an interior solution, this gives

$$\frac{G_t^{***}}{G_1^{***}} = \frac{\lambda^{***}r_t + \sum_i \alpha_i^{***}\partial(p_i^0/p_1^0)/\partial x_t^{***} + \delta^{***}\hat{p}_t^0}{\lambda^{***}r_1 + \sum_i \alpha_i^{***}\partial(p_i^0/p_1^0)/\partial x_1^{***} + \delta^{***}\hat{p}_1^0} \tag{11}$$

where *** denotes evaluation of the quantity concerned at this "third-best" optimum. Now, from the analysis of the previous section and our knowledge that $\hat{p}_t^0/\hat{p}_1^0 = r_t/r_1$, equation (11) reduces immediately to $G_t^{***}/G_1^{***} = r_t/r_1$.

Thus, in this case, the optimal shadow prices of traded commodities are their relative international prices, irrespective of whether there is a government budgetary constraint or not. Rule R_2 of Dasgupta and Stiglitz *is therefore misleading. The existence of a government budgetary constraint is* not *a sufficient reason for setting the shadow prices of traded commodities at values other than their relative international prices.* Two corrected versions of R_2 are that the relative shadow prices of traded commodities lie between their relative world and domestic prices: R_2', when there is a financial budgetary constraint on public projects expressed in domestic market prices; or R_2'', when there is a government budgetary constraint, and the domestic prices of those goods are distorted for reasons other than government tax policy (e.g., import quotas, or monopoly power exercised by import agents with exclusive import licences).

IV. Summary and Conclusions

This paper extends the case for shadow pricing traded commodities at their relative international prices in benefit-cost analysis. Assuming that any distortions in domestic market prices are due to nonoptimal tariffs and other taxes, and that project evaluators must take these distortions as given and fixed, we show that the correct shadow prices for traded commodities are their relative international prices (*a*) when there are nontraded commodities whose distorted prices are indirectly affected by public production of traded commodities, and (*b*) when there is a government budgetary constraint. Contrary to arguments found in the literature, neither of these cases in itself provides an argument for shadow pricing traded commodities at values other than their relative international prices.

References

Boadway, R. "Benefit-Cost Shadow Pricing in Open Economies: An Alternative Approach." *J.P.E.* 83, no. 2 (April 1975): 419–30.

Dasgupta, P.; Marglin, S. A.; and Sen, A. K. *Guidelines for Project Evaluation.* New York: United Nations, 1972.

Dasgupta, P., and Stiglitz, J. E. "Benefit-Cost Analysis and Trade Policies." *J.P.E.* 82, no. 1 (January/February 1974): 1–33.

Little, I. M. D., and Mirrlees, J. A. *Manual of Industrial Project Analysis in Developing Countries.* Paris: Org. Econ. Cooperation and Development, 1969.

———. *Project Appraisal and Planning for Developing Countries.* New York: Basic, 1974.

Rudra, A. "Use of Shadow Prices in Project Evaluation." *Indian Econ. Rev.*, n.s. 7 (April 1972): 1–15.

Sen, A. K. "Control Areas and Accounting Prices: An Approach to Economic Evaluation." *Econ. J.* 82 (March 1972): 486–501.

Weckstein, R. S. "Shadow Prices and Project Evaluation in Less-developed Countries." *Econ. Development and Cultural Change* 20 (April 1972): 474–94.

Miscellany

An Example of a Trading Economy with Three Competitive Equilibria

This brief note is presented merely as a convenience for those who wish to see what an actual numerical example of a smooth trading economy with multiple equilibria looks like when depicted in an Edgeworth box diagram. We present a two-trader two-commodity economy in terms of a fanciful exchange between two kinds of money. The example is robust, in that its qualitative features would survive small perturbations in the data.

The tourists.—Ivan has R 40 in his pocket and wants some dollars; John has \$50 to spare and would be happy to exchange some of it for rubles. Their utility functions (x = rubles, y = dollars) are

$$\begin{cases} u^1(x, y) = x + 100(1 - e^{-y/10}) & \text{(Ivan, in rubles)} \\ u^2(x, y) = y + 110(1 - e^{-x/10}) & \text{(John, in dollars).} \end{cases} \tag{1}$$

Note that these functions are not only concave and smooth (C^∞) but additively separable, with one good entering linearly in each. It is well known (but virtually ignored in many textbook treatments of competitive uniqueness)[1] that the competitive equilibrium is unique if the *same* good is linear and separable in all utility functions, provided only that this good is in sufficient supply and the preference sets are smooth (C^1) and strictly convex, as they are here. The present example shows that this "transferable utility" or "welfare maximization" approach to uniqueness does not allow even a modest tinkering with the hypotheses.

In figure 1 the indifference curves are indicated by the broken lines. The locus of points of tangency is the straight line D^1D^2, given by

$$y = x + 50 - 10 \log 110 = x + 2.995. \tag{2}$$

Work supported by NSF grant SOC71-03779 AO2 and ONR contract N00014-76-C-0085, respectively.

[1] See, e.g., Arrow and Hahn 1971, chap. 9.

[*Journal of Political Economy*, 1977, vol. 85, no. 4]
© 1977 by The University of Chicago. All rights reserved.

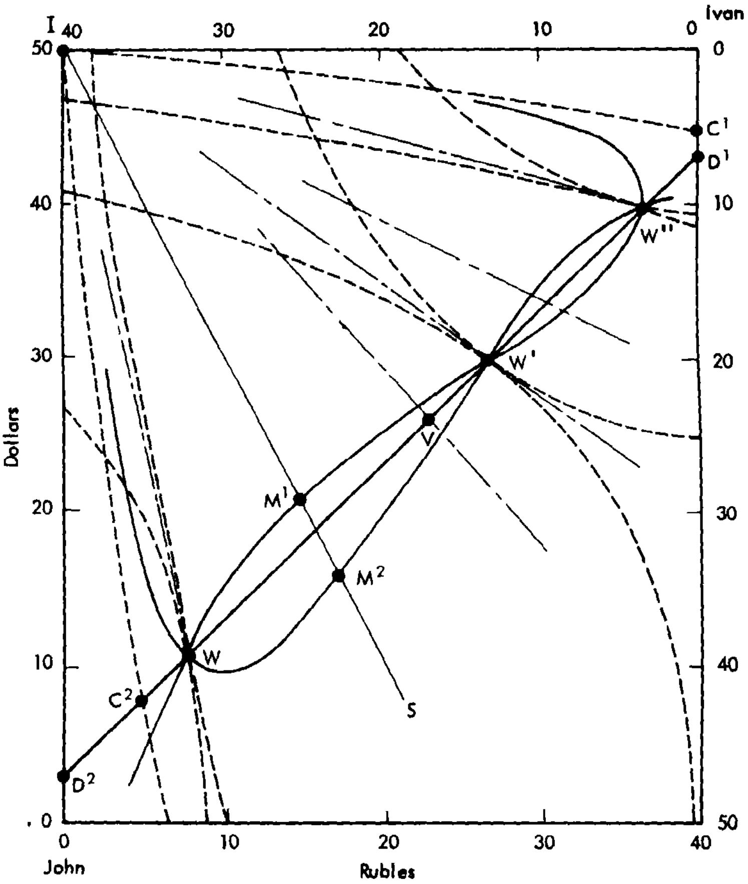


FIG. 1.—Three competitive equilibria

Edgeworth's "contract curve" C^1C^2 runs along this line and a short piece of the boundary. The conditions for a competitive allocation reduce by elementary calculus to the following transcendental equation:

$$x(1 + 11e^{-x/10}) = 10 \log 110, \tag{3}$$

which has three roots in the region of interest. These lead to the three solutions indicated by W, W', W'' in figure 1 and given numerically in table 1. Their relation to the two response curves (solid lines) is also shown in figure 1.[2]

---

[2] To illustrate the definition of "response curve," suppose the price ray IS is given exogenously. Then Ivan's best trade is M^1, John's M^2.

TABLE 1

NUMERICAL DATA FOR FIGURE 1

	ALLOCATION (TO JOHN)*		EXCHANGE RATIO	UTILITY PAYOFFS		
	Rubles	Dollars	(Dollars: Rubles)	Ivan	John	
I	0.00	50.00	...	40.00	50.00	Initial point
C^1	40.00	44.89	0.20:1 (m) 0.13:1 (a)	40.00	152.88	Endpoints of core
C^2	4.83	7.83	6.79:1 (m) 8.73:1 (a)	133.69	50.00	
W	7.74	10.74	5.07:1	130.29	70.01	Competitive solutions
W′	26.83	29.82	0.75:1	99.88	132.30	
W″	36.78	39.77	0.28:1	67.27	146.99	
V	23.00	25.99	1.10:1 (m) 1.04:1 (a)	107.94	124.96	Game value

NOTE —(m) = marginal, (a) = average.
* For Ivan, subtract from (40, 50)

If we take a contract point between *W* and *W′*, the direction of common tangency (dot-dash line) passes above the initial point *I*; if we take one between *W′* and *W″*, it passes below *I*. This means that the equilibrium prices associated with *W′* are dynamically unstable, in the sense that raising the price of either good would create a positive excess demand for that good. This in turn (in a suitable dynamic model) would tend to drive that price up still further. The two other solutions, *W* and *W″*, are dynamically stable (see Gale's article [1963], where another simple example of nonuniqueness will be found).

In table 1 we have also indicated the core and value solutions of the trading game in order to suggest outcomes alternative to those of the competitive equilibrium (see e.g., Shapley and Shubik 1969).

L. S. SHAPLEY

RAND Corporation

M. SHUBIK

Yale University

References

Arrow, K. J., and Hahn, F. H. *General Competitive Analysis.* San Francisco: Holden-Day, 1971.

Gale, D. "A Note on Global Instability of Competitive Equilibrium." *Naval Res. Logistics Q.* 10 (March 1963): 81–87.

Shapley, L. S., and Shubik, M. "Pure Competition, Coalitional Power, and Fair Division." *Internat. Econ. Rev.* 10 (October 1969): 337–62.

Editors' Notice

In the February 1975 issue of the *Journal of Political Economy* we published a paper by Dr. Okonkwo A. Nwani on "The Quantity Theory in the Early Monetary System of West Africa with Particular Emphasis on Nigeria, 1850-1895." This paper made extensive use of two articles by Marion Johnson, but that use was neither adequately documented nor accurate in its reading. We apologize to Mrs. Johnson and request our readers who are interested in the subject to read her articles, "The Cowrie Currencies of West Africa," *Journal of African History* 11, nos. 1 and 3 (1970): 17-49, 331-53.

Book Reviews

Foreign Trade Regimes and Economic Development: India. By JAGDISH N. BHAGWATI and T. N. SRINIVASAN. A Special Conference Series on Foreign Trade Regimes and Economic Development, vol. 6.
New York: National Bureau of Economic Research (distributed by Columbia University Press), 1975. Pp. 261. $15.00 (cloth); $5.00 (paper).

The National Bureau of Economic Research, with the financial support of the Agency for International Development, has sponsored an ambitious research project on the trade regimes of developing countries, including Brazil, Chile, Colombia, Egypt, Ghana, India, Israel, Philippines, South Korea, and Turkey. The book under review by Bhagwati and Srinivasan is the sixth volume in the country studies.

A similar exercise was performed under the auspices of the OECD some years ago. The study on India, entitled *India: Planning for Industrialization*, was done by Bhagwati and Padma Desai, and the book under review by Bhagwati and Srinivasan freely draws on the earlier volume and may be regarded as its sequel.

This study consists of four parts. The first part presents an overview of the economic developments in India from 1950 to 1970. From 1951 to 1966 there were three successive 5-year plans; a fourth was launched in 1969. Based on a Harrod-Domar framework, the First Plan (1951–56) attempted to build the infrastructure necessary for growth and to raise the domestic saving rate of the economy. The Second Plan, inspired by the Feldman-Mahalanobis model, was bolder in scope and conception and attempted to influence the pattern of investment and not merely its volume. It created a base of capital-goods industries largely under the direction of the public sector to enable the private sector to produce domestically the consumer goods that were demanded in the economy. The fundamental structural changes required by import substitution inevitably led to imbalances of various kinds resulting in public control over domestic production and foreign trade. This took two main forms: (1) industrial targeting and licensing and (2) exchange control and licensing of imports. Provoked by a serious balance-of-payments crisis in early 1957, a regime of quantitative restrictions (QR) on international transactions was imposed and continued in an intensified form up to 1962.

From 1962 to 1966 significant steps were taken to soften or modify the impact of this regime—export subsidization was started in 1962 and was intensified during the period on a wide range of goods, and the government deployed import duties to offset the undue advantage that an overvalued exchange rate conferred on the importer. These moves culminated in the 1966 devaluation and the import liberalization measures such as reduction of import duties and elimination of export subsidies that accompanied it. In the period 1966–68 the liberalization experiment ran into some heavy weather, and during 1968–70 there was a return to the QR regime that existed prior to the devaluation of 1966.

During the period 1950–60 price increases were remarkably moderate, averaging about 3 percent per year. The Sino-Indian border war of 1962, the Indo-Pakistan war of 1965, the droughts of 1965–66 and 1966–67, and the war with Pakistan in 1971 that led to the creation of Bangladesh were major exogenous shocks to the Indian economy that adversely affected both the growth rate and the stability of prices that prevailed earlier. During the 2 years preceding the Declaration of Emergency in June 1975 the consumer price index was increasing at an annual rate in excess of 20 percent.

In part 2, the authors make a detailed analysis of the exchange control regime that existed in India during the period 1956–66 and attempt to determine its economic effects. They marshall a formidable list of arguments against the policy of industrial licensing and import control and lay much of the blame on the inward-looking bias of the Indian planning strategy. This is followed by an examination of export policy and performance during this period, and the conclusion is reached that export promotion policies were inefficiently designed and implemented.

The "liberalization episode" of 1966–68 is considered in part 3, and the conclusion reached that the adverse changes in the price level, production, and investment were not due to the June 1966 devaluation and other liberalization policies but were largely due to forces that emanated from the droughts of 1965–66 and 1966–67. The authors also challenge the view widely held by large developing countries that the price inelasticity of export supply and/or demand makes devaluation disadvantageous to their interests. They draw the lesson from the Indian experience that the timing of a devaluation cum liberalization is of decisive importance and that it should come just after a good harvest, so that a "smart policy-maker could use the improvement in export performance thanks to a good harvest to advantage by crediting the devaluation with this success"!

This is perhaps the best section in the whole book. Eschewing a doctrinaire stand, the authors by means of empirical and logical reasoning present a balanced and convincing assessment of the "liberalization episode." The devaluation of June 1966 was accompanied by a substantial elimination of export subsidies and reduction of import duties in the hope that it would be a more efficient substitute for these measures. In any analysis of its effects, therefore, the distinction between gross and net devaluation assumes significance. The total liberalization package could be expected to have only a negligible impact on the export of traditional exports and no more than a mildly expansionary impact on nontraditional exports. Its effectiveness could also be reflected in a decline in uneconomic exports. Unfortunately, such sophisticated understanding was lacking, and events exogenous to the liberalization package further undermined its beneficial impact. The fact that devaluation was undertaken in response to "aid blackmail" by Western donors also contributed to its political unpopularity. By 1969–70 most of the liberalization measures had been abandoned, export subsidies were reinstated, and the QR regime was firmly in the saddle. In this sense, the authors concede that the liberalization episode had failed. But they are able to demonstrate that factors external to the liberalization package were responsible for this failure. Therefore, their conclusion that the improvement of India's export performance requires the deployment of the tools of exchange-rate policy and a more efficient and less wasteful form of domestic protection seems to be logically unassailable.

It is when they pass from static efficiency effects to growth effects of foreign trade in part 4 that their empirical evidence and logical argument become less convincing and more open to question. With admirable scholarly candor they make explicit the limitations of the methodology they employ, as when using the Eckaus-Parikh model, and generally show an awareness that most of the engines of analysis they employ in this section are capable of reaching only mildly suggestive conclusions.

Throughout the book an unresolved ambivalence seems to hover around what the authors wish to accomplish. By and large, the earlier sections draw attention to the wastes and inefficiencies in India's foreign trade regime and persuasively argue the case for the discipline of the market through such measures as trade liberalization, selective delicensing, and, when necessary, exchange-rate adjustment. These are policy instruments of singular potency, and their judicious use can greatly improve India's economic performance and trade. To have given a convincing demonstration of the urgent need for such a reform in India is in itself a worthy accomplishment. But it becomes evident in part 4 that the authors are after bigger game. There the shafts are aimed not only at the policy instruments, but at the whole strategy of development through import substitution. This effort has not been as successful. Whatever may be the merits of export-led development strategy for small territories like Hong Kong, Taiwan, and South Korea, for a country of continental dimensions like India, the import-substitution strategy may be the only realistic choice.

The challenge facing the Indian planner is to implement this strategy with the help of policy instruments that reduce waste and increase efficiency. Unimaginative and cumbersome controls and bureaucratic rigidities and corruption are as damaging to "learning by doing" as total absence of protection. In devising an efficient and less wasteful form of protection and in improving the general performance of the economy the market mechanism can be a powerful aid. But the sheer size of the economy and the fact that exports account for less than 5 percent of her national income preclude reliance on export-led growth in the Indian case. Many of the problems that bedevil the Indian economy, including those related to foreign trade, are due to India's failure to carry out the fundamental and far-reaching reforms in her domestic economy necessary to bring about sustained increase in production and restore balance in her savings-investment account through increased domestic resource mobilization. Not only is it unrealistic to expect the international trade tail to wag the domestic economy dog in the Indian case, but a shift to an export-led growth strategy is also incompatible with other goals of Indian planning, such as self-reliance.

To dispute the conclusions of the book, however, is not to question the subtlety and rigor of its analysis. As is only to be expected when two eminent scholars join forces, the resultant product contains much that is interesting, stimulating, and valuable. Both the policymaker and the development economist have much to learn from it.

KOILPILLAI J. CHARLES

Lakehead University

American Farm Policy, 1948–1973. By WILLARD W. COCHRANE and MARY E. RYAN. Minneapolis: University of Minnesota Press, 1976. Pp. 431. $18.50.

Cochrane and Ryan have undoubtedly put together the most complete available chronology of farm commodity programs during the postwar years.

Credentials of the authors are impressive. Willard Cochrane is professor of agriculture and applied economics at the University of Minnesota and is the author of a number of books, including *Farm Prices: Myth and Reality*, and *The City Man's Guide to the Farm Problem*. He achieved his greatest prominence as the architect of farm commodity policy during the Kennedy-Freeman years of the early 1960s. Mary E. Ryan is a research fellow in agricultural and applied economics at the University of Minnesota and has written on other agricultural subjects. Financial support for the project was provided by the Agricultural Experiment Station of the University of Minnesota.

The authors indicate that the book is basically a reference, which it is. There is thorough and accurate compilation of legislative actions and administrative decisions taken during the 25-year time span under examination. But there are also interpretations drawn from the experience cited.

The farm commodity programs under review in this book are the price supports, nonrecourse loans, acreage controls, marketing quotas, government payments, stock accumulation, and surplus disposal that characterized the quarter-century under review.

The book is divided into three parts. Part 1, which consists of four chapters, presents the background and setting. Part 2, also four chapters in length, is the heart of the book. It contains all anyone is likely to want regarding the detailed operation of farm commodity programs. This material is presented fully and fairly. Part 3, the briefest section, offers two chapters addressed to consequences and appraisal.

Since part 3 will be of greatest interest to most readers, this review will be concentrated thereon.

The authors select five categories within which to examine the consequences of the program.

1. Farm prices and incomes. The authors conclude that farm prices and farm income were higher from 1952 to 1972, as a result of the operation of government programs, than they would have been in the absence of these programs. How much higher they find it impossible to say.

2. Distribution of producer incomes and wealth. Farm programs were adopted to help the commercial farmers, who had something to sell. The authors state that protagonists of farm programs may have misrepresented commodity programs as being necessary to help small-scale farmers.

3. Resource allocation, efficiency, and output. The writers say that 5–10 percent more resources were committed to agriculture during this period than were in fact needed. Programs of price support, direct payments, production control, surplus disposal, and stock accumulation "did not, we argue, create the excess capacity, or resource allocation problem; it came about in response to the resource allocation problem." They say that no effort was made to reallocate these resources to some other use. They recognize that the price and income insurance provided by the programs led to a rapid advance in farm technology: "The farm policy compromise in conjunction with rapid and widespread technology advance, operating in the context of an atomistically competitive industry, produced the important result—chronic food abundance."

4. Foreign trade. The writers conclude that adoption of a free-trade policy throughout the noncommunist world would have had a moderate but noticeable effect on the location of agricultural production, would have increased rather importantly trade in agricultural products involving the United States, and would have increased the real incomes of consumers in the United States by some modest amount. They conclude that with the widespread interference with markets for farm products no one country could significantly expand its trade through the unilateral adoption of free-trade policies.

5. Consumer welfare. "The price of food to consumers in the United States would, without question, have declined in almost any given year from 1948 to

1965 if the farm and income support program had been eliminated for that year."

Missing from this list of categories is an estimate of the cost to the treasury of these massive programs. Also missing is awareness of the loss, on the part of farm people, of decision-making prerogatives which occurs when pricing and production policies are determined by government. The tendency in this book is to treat decision making as neutral, having no utility of its own, being willingly surrendered to government if an extra dollar could be obtained thereby. The evidence is contrary to that assumption.

Cochrane and Ryan distill from their examination these 10 principles to guide future farm policy architects: (1) The programs must be highly flexible. (2) Farm commodity prices should be stabilized at, near, or approaching the world level. (3) A major reserve stock program is essential. (4) A reserve stock program needs (*a*) an export policy to monitor and regulate the flow of commodities from the United States and (*b*) a supply management policy. (5) Programs of supply management should move in the direction of negotiable marketing quotas. (6) Income assistance should be in the form of direct payments. (7) Domestic food-assistance programs should be based on needs of the consumers. (8) Foreign food aid should cease to be a residual claimant on U.S. food production; it should be based in large part on need. (9) Attention should be given to assisting hired farm workers and low-production farmers. (10) Supply management policies should be symmetrical, so that if a prolonged period of short food supplies should arise production could be stimulated.

Those who know Willard Cochrane well will note in this book strands of his long-held views favoring government intervention in the pricing and production of farm products. They will recognize also some modification of these views.

The book deals primarily with the old farm policy agenda—commodity programs. It does not deal in any depth with new farm policy subjects—environmental issues, natural resource use, rural development, international agricultural matters, and the like. To call attention to this fact is not to fault the book; it does not purport to cover the entire range of issues. I make this point only to underscore the fact that farm policy issues of the future will not be limited to commodity programs.

The book ends, really, with the Agricultural Act of 1973. At that time profound changes were underway, changes in supply, demand, public attitudes, and political power.

The possibility that the future might differ distinctly from the past in several respects raises a question as to the applicability of conclusions drawn from the experience of the 25-year period under examination.

In any case, we have here a good reference as to what actually happened in the past. The authors provide us with their interpretation of these events; other students are free to supply their own.

The book would be a useful reference for a course in agricultural policy. It would be good collateral reading or a good case history for a course in government. And to anyone who has been a participant in the postwar farm policy battles, it would evoke happy memories or recall moments of frustration—perhaps both.

Don Paarlberg

Purdue University

Oil Resources: Who Gets What How? By KENNETH W. DAM.
Chicago: University of Chicago Press, 1976. Pp. 193. $11.95.

For a number of years Kenneth Dam has been publishing original work on the policies of the British and Norwegian governments toward the exploitation of North Sea oil and natural gas. His research, mainly appearing in the *Journal of Law and Economics*, as well as embodying the degree of detachment one would expect from across the Atlantic is of sufficient importance that it should have caused some fundamental rethinking by the European civil servants and polititians who formulate North Sea policy. That there is little sign of any significant redirection of policy is no fault of Dam: it is more an indication of the strength of self-interest among the bureaucracy which leads to the continued resort to administrators' "discretion" rather than to the increased use of the market for which Dam so cogently argues. In this admirable book much of his research on North Sea policy which has previously appeared is integrated with further material, including some brief remarks on United States's experience with competitive bidding for federal oil and gas leases and on company-government relationships in the OPEC countries, and the book is up-to-date enough to include comments on the first British "participation" agreement announced in February 1976.

The core of the book is a case for allocating petroleum licenses by competitive bidding (auctions) rather than by administrative discretion. Petroleum resources are generally initially in the ownership of government and, as a consequence, some positive government decision is necessary on how exploitation should proceed. Both the British and the Norwegian governments have opted for the discretionary system, apart from a limited experimental auction held by the British government as part of the fourth round of license allocation in 1971. Indeed, the United States appears to be the only major country in which the auctioning of petroleum licenses is common.

Unwillingness to auction potentially valuable resources seems curious at first sight. As Dam points out, competitive bidding in the absence of collusion should lead to the state's receiving a large part of any economic rent which accrues from development of the resources. In the early days of exploration and development of any new petroleum province, imperfect knowledge will make it difficult for bidders to know what value to place on the available land. However, no one should be looking for perfect solutions in an imperfect world, and the auction seems a good if rough-and-ready method of extracting rent as knowledge accumulates. Auctioning licenses solves the allocation problem, collects the rent, and should favor low-cost potential licensees who can afford to outbid competitors. The popular objections to auctions—for example, that they collect money which would have been better employed in drilling for oil—are examined by Dam and all are found wanting. There is no evidence that bonus payments for licenses detract from drilling programs. In fact one would have thought the reverse is more likely to be true, since a company with a large negative entry at the beginning of a project's cash flow has a powerful incentive to prove oil or gas and bring in revenue rapidly, as compared with a situation in which rent is collected throughout the project life via the tax system. Only if governments collect substantially less rent through taxation than they would have received from competitive bidding might drilling be stimulated, but, as Dam says, it would then presumably rise above the optimum. In general Dam supports bonus rather than royalty bidding because of the discouraging effect of high royalties on the development of marginal fields and their tendency to cause early abandonment of fields.

The British system of allocating licenses in the 1960s offers perhaps the most tangible evidence of the effects of discretionary allocation since such effects can now be seen over a fairly long period. Dam analyses them in some detail. Chapter 8 provides what, to one who was involved in the 1966–68 negotiations between the oil companies and the (then) Gas Council, seems a remarkably accurate account of the main events. The problem which faced the British government in those years, after some substantial natural gas discoveries in the southern British North Sea, was that the production licenses had in effect been given away. They were "licenses to print money," to use a British phrase hallowed in another context. The licenses were given to companies thought worthy by the (then) Ministry of Power after a complex process of negotiation over work programs and other matters. Whether no one realized that the oil companies would be left with the rent or whether, as seems more likely, the officials wanted to keep close control of the negotiations to sell the gas is now immaterial. The costs of not auctioning then began to show.

The oil company/Gas Council discussions, though inevitably complex, might have proceeded fairly smoothly if the rent issue had been out of the way. As it was, however, there was continual interference from the Ministry of Power—which was alarmed, in a way only civil servants working for a Labor minister can be, at the prospect of the oil companies making money from the North Sea—and a protracted debate began, not about the price the Gas Council should pay for natural gas but about the theory behind price setting. The ensuing discussions, over 2½ years, about whether a market-related or a cost-plus basis for price should be used were essentially a waste of time rendered necessary only because the licenses had been issued for free.

In the circumstances, the officials concerned seem to have decided that the Gas Council would have to be used as the rent collector by strict application of the 1964 Continental Shelf Act so as to deny to the oil companies any opportunity of selling natural gas to anyone other than the Gas Council (apart from the relatively minor exception of nonfuel uses). The natural gas prices eventually set in 1968–69 for sales by the producing companies to the Gas Council were low in relation to the fuel prices then ruling and very low indeed compared with the higher fuel prices of the early and mid 1970s.

This set off a chain of further problems. If rent is available and should be removed from the companies—a point about which there is some scope for argument—it should presumably accrue to the state. There can be no case for channeling it from one large corporation which happens to be privately owned to another large corporation which happens to be "nationalized." But if the rent accrues initially to the nationalized corporation, as was the case in Britain, how is one to know that some or most of it does not remain there? To the naive, it might appear that because of the low natural gas purchase price the Gas Council would sell at market prices and make large profits which would then be taxed away by the government. Those more knowledgeable about large organizations would, however, guess that the distribution costs of the Gas Council would be unlikely to be as low as they would have been at a higher purchase price for natural gas and that what might have been rent in private corporations would largely be translated into organizational slack in a nationalized corporation. This hardly seems desirable, and it could have been avoided by auctioning production licenses in the first place. To some extent consumers may have obtained a short-term benefit from the low price which the Gas Council paid for natural gas because it showed up in sales by the council at somewhat less than the market price but, predictably, gas sales to industry had eventually to be rationed because of excess demand. Furthermore, as Dam shows, there was probably some reduction in exploration effort as a result of the low price paid to the producing companies. There still remains a fascinating study to be done of where the rent disappeared, but Dam has provided a valuable commentary on the complications which arose from the decision not to auction licenses, which is an essential foundation for those who wish to work on the subject.

There are several points where one could argue with the opinions expressed in the book. For example, the treatment of the Norwegian and British oil taxation systems is rather brief and does not properly convey the differences between the two systems: the impression is given (p. 69) that the British system is more favorable to marginal fields whereas, in fact, the Norwegian system seems to contain a bigger tax differential between small and large fields (Morgan and Robinson 1976). Similarly, there are only a few remarks about depletion policy which is closely related to license allocation (Robinson and Morgan 1976). More generally, belief in the idea of state rent collection in the ultimate depends on one's conception of the function of government. Not everyone would agree that rent should be collected from one section of the community but not from others nor that one can assume wiser use of the rent by government than by private companies. Nevertheless, Dam's general case for the auction system as a device for rent extraction superior to monopoly purchasing agencies, royalties, taxation, and government participation is extremely powerful.

Indeed, it is sufficiently powerful for one to ask why it has not already had more impact, especially since in Britain the experimental auction of 1971 was successful in terms of yielding substantial revenues. Dam points out that the discretionary system of allocation in Britain has been used for nationalistic reasons and for informal arm twisting of the oil companies. Clearly his ideas could be expanded, using the concepts of the economic theory of democracy. If there is one thing above all which characterizes the discretionary system of license allocation and distinguishes it from competitive bidding, it is the amount of power and influence it gives to the officials who administer it. Consequently, they believe they have a strong interest in its survival. One does not need to think of them as in any way malevolent but merely as pursuers of their personal interests, in the same way that the private sector is supposed to be self-seeking but without the same desirable results. The British and Norwegian governments have gradually taken ever closer and more detailed control of North Sea activities in recent years. Many of us believe, like Dam, that less regulation of North Sea activities is highly desirable. We now need someone to produce a means of persuading those whose employment and advancement prospects depend on detailed intervention in industry that some withdrawal of the services which they lavish on us would be one of the greatest benefits they could bestow on the community. It seems a task ideally suited to the persuasive powers of Chicago!

Colin Robinson

University of Surrey

References

Morgan, Jon, and Robinson, Colin. "Comparative Effects of the U.K. and Norwegian Oil Taxation Systems on Profitability and Government Revenue." *Accounting and Bus. Res.* 6, no. 25 (Winter 1976): 2–16.

Robinson, Colin, and Morgan, Jon. *Economic Consequences of Controlling the Depletion of North Sea Oil and Gas*. London: Trade Policy Res. Centre, 1976.

Housing Markets and Racial Discrimination. By JOHN F. KAIN AND JOHN M. QUIGLEY. New York: National Bureau of Economic Research, 1975. Pp. 393. $17.50.

As the third in a series of books from the National Bureau, this book continues the study of housing submarkets in metropolitan areas by analyzing particularly detailed housing information for St. Louis in 1967. The two previous studies were on Detroit (Ingram, Kain, and Ginn 1972) and San Francisco (Straszheim 1975).

After an introductory chapter, the theory guiding their work is outlined in chapters 2 and 3. Chapter 4 discusses the St. Louis data in detail. Chapters 5 and 6 analyze the probability of home ownership and are only tangentially related to the theory. A traditional analysis of housing expenditures is covered in chapter 7. The analysis of housing attributes is covered in the next three chapters. Chapter 8 develops a sort of hedonic index of the valuation of the attributes of housing. Chapter 9 establishes demand relations for the various attributes, and chapter 10 is an attempt to investigate the jointness of demand for attributes of housing. The final chapter, 11, is a summary of the findings.

The main theme of the book can be stated in the following way. Housing is not a homogeneous good in continuous long-run equilibrium. Rather, housing is better characterized as a bundle of attributes. Furthermore, since housing is a durable good and since some of the attributes are not reproducible, the price of housing includes costs of production and quasi rents. Prediction of values, then, will depend upon any systematic relations between socioeconomic characteristics of households and housing attributes and will require empirical analysis. The gross price of housing includes production costs, quasi rents, and transportation costs (explicit and implicit). Families choose their housing by selecting the bundle with the desired characteristics with the minimum gross price. In the empirical analysis, transportation costs are ignored. Blacks, however, do not have the same choices as others. The housing market is segregated into two submarkets: one for blacks and one for whites. This restriction in supply leads to a number of inequities. Blacks do not have the same opportunity for homeownership, which restricts their welfare since they cannot take advantage of the tax breaks for homeowners and since they cannot receive the appreciation of value that has occurred in general in the 1950s and 60s. Blacks' ability to own is restricted by the supply of single detached dwellings in ghetto areas and by the higher price for given quality that they are required to pay. The interrelated analysis of expenditures and attributes shows blacks to have paid more for a given quality but to have spent less in total for housing than whites with comparable socioeconomic characteristics.

The analysis of discrimination unfortunately is marred by some emphases that are inconsistent with the empirical evidence that is admirably summarized. Kain and Quigley appropriately begin with an assumption of segregated housing submarkets. While boundaries may be blurred today, during the 50s and 60s they were not. Instead of making market separation their conceptual framework, however, they concentrate on whether blacks pay more for a given quality of housing than whites. The myopia may be a result of concentration on the cross-section data for 1967. Other studies cited by the authors, however, have shown that there is more than one price comparison for black and white occupancy and that these comparisons vary over time and with different circumstances.

A theory about why blacks pay more should be able to explain why the differential exists and how it changes over time. Kain and Quigley rely on discrimination and the resulting supply restrictions. A more complete explanation is to rely on Bailey's original story (Bailey 1959). This story also begins with the idea that the markets are separated between black and white occupancy. Within each of the segregated markets there are separate prices for housing, but there are also different prices along the boundary between the markets. Whites pay low prices on the boundary or in areas where change is anticipated (Little, Nourse, and Phares 1975). Blacks pay a boundary premium over ghetto prices to escape. So long as the black boundary price is greater than that for whites (taking account of conversion costs), the ghetto expands.

During the 50s and 60s, rapid migration of blacks to cities, natural increase in black population, and destruction of housing for black families by road construction and urban renewal caused aggravated excess demand in ghetto areas. Ghetto areas expanded as a result of the upward pressure on prices within ghettos and on the black side of the boundary making conversion of houses along the boundary profitable. In the late 60s and 70s, the migration of blacks to cities declined, and there was less excess demand. Blacks were still trying to escape the ghetto so that they were still willing to pay a premium for boundary property over that within the ghetto. Whites, on the other hand, were still discounting housing in those areas in which transition was anticipated, so that profitable arbitrage continued to increase the ghetto. Within the ghetto, however, there was an excess supply of housing, so that prices declined, vacancies rose, and houses were abandoned, vandalized, and destroyed. In fact, those black families escaping the ghetto in the late 1960s by acquiring housing under liberalized FHA guidelines were caught in the middle as the ghetto boundary moved to other areas and their houses became part of the ghetto.

Besides diverting the authors from an analysis of change, the concentration on a comparison of black and white payments for a given quality of housing leads them to misinterpret their regressions. For example, in chapter 8 housing expenditures were regressed against the vector of attributes to determine prices or value for each attribute—a hedonic price index. Although the range of attributes was greater than in previous studies, the sample was heavily biased toward city houses. According to their equations, increases in the proportion of whites in a neighborhood causes expenditures to be less. Kain and Quigley attribute this to the so-called markup for housing for blacks. Given that the sample is dominated in nonghetto areas by city housing, their regression really compares expenditures for all housing occupied by blacks with those by whites in the areas discounted in anticipation of change. This is only one of the possible comparisons (not even the most useful), and with changing circumstances this differential should change from what it was in 1967.

A major element in the Kain-Quigley analysis is the restriction of housing supply to blacks. Their evidence is indirect and relies on avoidance of internal conflicts within their logic and evidence. For example, a major discussion of the problem appears in the analysis of homeownership. They run regressions of the probability of homeownership against a number of reasonable independent variables depicting age, life-cycle stage, income, and stability of income. Their results suggest that blacks have significantly lower probability of homeownership. Three possible hypotheses are considered: blacks have a different taste; blacks have less wealth; or blacks have less access to housing because of supply restrictions or price discrimination. Taste differences are rejected because separation of equations by race shows no significant differences in the independent variable depicting tastes. Although lower wealth may be a contributing factor, it is difficult to prove since homeownership is a significant proportion of the wealth of low- and middle-income families. Thus, one is left with the hypothesis that the differential is a result of supply restrictions.

Further corroboration of the hypothesis is found in the analysis of housing attributes. Each of the housing attributes was regressed against the same income, age, and life-cycle variables used in the ownership study. The estimates show that blacks have the same demand characteristics as whites for some attributes, such as number of rooms, but that blacks have large differences from whites in demand for such attributes as newness and lot size. The differences were attributed to the likely restriction from black occupancy of housing with the latter characteristics.

The last link in their story is that blacks spend less on housing because there is price discrimination against them. Abandoning the entire conceptual framework that housing is a bundle of attributes rather than a homogeneous good, Kain and Quigley find that housing expenditures by blacks were less than whites for comparable age, life cycle, and tenure status. At the same time, their attribute analysis, which is consistent with their theory, showed that quality cost more inside than outside the ghetto (subject to our previous objection). The authors argue that blacks spend less because the price per unit of a given quality house is higher to them. The reason is that the price elasticity of demand for housing is greater than one, a fact borrowed from previous studies and not estimated in this one. But what is the meaning of a price elasticity of demand for a nonhomogenous bundle of attributes? The analysis is inconsistent with their initial assumption that housing is a bundle of attributes.

Given the wealth of detail and analysis, so much of which could not be discussed in a short review, this book is, surprisingly, a disappointment. The reason is that the theory is written as if the black-white differential is timeless. We are not shown how the segregated-market model will change internal and boundary prices with changes in circumstances, so that the model is not useful in evaluating the impact of alternative public policy proposals on the market. The reduction in migration of blacks to central cities, the decline in total SMSA population in some areas, such as St. Louis, and the relative increase in new housing costs would change the differential and operation of the markets significantly, even if segregated markets were as strong as ever and had not been blurred by changes in attitudes since 1967.

HUGH O. NOURSE

University of Missouri —St. Louis
and
Hoover Institution on War, Revolution and Peace

References

Bailey, Martin J. "Note on the Economics of Residential Zoning and Urban Renewal." *Land Econ.* 35 (August 1959): 288–92.

Ingram, Gregory K.; Kain, John F.; and Ginn, J. Royce. *The Detroit Prototype of the NBER Urban Simulation Model.* New York: Nat. Bur. Econ. Res., 1972.

Little, James T.; Nourse, Hugh O.; and Phares, Donald. *The Neighborhood Succession Process.* HUD-PDR-104. Washington: Department of Housing and Urban Development, July 1975.

Straszheim, Mahlon R. *An Econometric Analysis of the Urban Housing Market.* New York: Nat. Bur. Econ. Res., 1975.

Manufactured Export Expansion and Industrialization in Brazil. By WILLIAM G. TYLER. Kieler Studien 134.
Tübingen: J. C. B. Mohr, 1976. Pp. 373.

This book is composed of nine chapters, five technical appendices, and a statistical appendix. Its author states two objectives in the preface: (*a*) "an improved basis for comparison with other countries" and (*b*) a better "understanding of Brazil's specific experience." In neither respect does the book improve on the existing literature. This review discusses some of the questionable points only; little or no attention is paid to those portions of the book that contain correct analyses since most of these have appeared elsewhere.

Chapter 2 presents a general overview of the industrialization and growth of Brazil since the last century. This chapter, essentially a descriptive summary, is sprinkled with unhelpful and largely speculative comments on past governmental policies. The statement that Brazilian growth during the period 1968–73 "can largely be attributed to efficacious economic policy within the capitalistic framework" is an overly strong statement, inasmuch as growth was widespread throughout the world, and during this period there was a substantial increase in governmental intervention in Brazil.

Chapter 3 attempts to quantify the importance of the import substitution process and of export expansion in contributing to final demand growth. Following Chenery (1960), the identity Δ total demand $= \Delta$ domestic demand $+ \Delta$ exports $- \Delta$ imports is used to analyze the relative importance of each demand component for 21 industrial sectors. To take into account direct and indirect effects, the 1971 input-output table for Brazil is used as suggested by Morley and Smith (1970). The 1971 table is used to analyze the periods 1949–64, 1964–67, and 1967–71. The result is as expected: the internal demand was the most relevant component of the total demand in all periods. The fraction of total demand generated by the import substitution process was about 24 percent in the period 1949–64, and negative after 1964 due to import liberalization (according to Tyler). It is interesting to note that, although the share due to exports was minimal, it grew from 2 percent in the first period to 6 percent in the last. The exercise was repeated to include the secondary income effects, using a Keynesian multiplier approach. Although interesting, this is probably misleading since Keynesian analysis refers to the short run and assumes unemployment, whereas the Brazilian economy appears to have been working close to full capacity, at least during the final period.

To complete the analysis of Brazilian industrial growth, the supply side is discussed and the efficiency of the industrial sector examined (chapter 4). The main conclusion is that Brazilian industry is reasonably competitive. Unfortunately, the evidence is weak. Although the difficulties of estimating production functions are stated, CES functions are estimated for each of 21 industrial sectors using cross-section data for states and then used uncritically to draw inferences about economies of scale, technology, and efficiency.

Cobb-Douglas production functions are also estimated from a Visão sample of firms' balance sheets, using sales as proxy for product. Thus, $X_i = A_i L_i^{\alpha} K_i^{\beta}$ is actually $P_i X_i = P_i A_i L_i^{\alpha} K_i^{\beta} = A_i^{*} L_i^{\alpha} K_i^{\beta}$ for estimation purposes. The OLS are used to estimate this equation on its logarithimic form, and inferences are drawn from these regressions. Nerlove's classical work (1965) is cited in vain since economies of scale are discussed by comparing $\alpha + \beta$ to *one* and no attention is paid to the impossibility of having $\alpha + \beta = 1$ in a competitive market if L_i and K_i are endogenous and determined by profit-maximization behavior subject to the production function.

Using these Cobb-Douglas production functions, the author attempts to test the differences in technology used by foreign and domestic firms by comparing the constant of the regressions for the two groups of firms in each sector. Because observations are few this test is performed by means of a *dummy* variable, assuming value one for foreign and *zero* for domestic firms, to conserve degrees of freedom. The author concludes that for three sectors (machinery, drugs, and textiles) there is a difference in technology used by these two groups of firms, and also for industry as a whole when all data are considered together in the same regression, since the coefficients for the dummy variable are significantly different from zero only in those regressions.

But this is not an appropriate test due to the data used. Since the dependent variable is sales (i.e., P_iX_i), the constant is in fact log (P_iA_i), and therefore two effects are confounded in the constant: the price and the technical parameter A. Thus, the coefficient of the dummy variable might be significant due to price differences which, in fact, are due to quality differences; and, if quality varies, it does not make much sense to compare the technology since we are dealing with two different product mixes. In fact, the sectors for which the coefficients of the dummy variable appeared significant are those where quality differences are very important.

Section 2 of chapter 4 presents some additional material on industrial efficiency in Brazil. The indicators of industrial efficiency are unreliable, and some of the statements do not follow either from economic theory or from the existing evidence (e.g., "One of the reasons domestic prices are high is that profits are high" [pp. 107–8]). The comments on specific industries are at times misleading (e.g., textiles). Price comparisons are made without accounting for quality; indeed, this problem is not even mentioned, and theoretical statements which are valid under certain circumstances are applied to inappropriate situations.

Chapter 5 gives a description of Brazilian exports and export performance during the postwar period. Except for section 2 it mainly reproduces other people's findings. The effort to capture the effect of the Latin American Free Trade Area (LAFTA) on Brazilian export performance is unsuccessful for two reasons. One has to do with LAFTA itself; the author fails to consider the expansion of the number of participating countries in his analysis. The second difficulty has to do with the data used. It is probable that Brazilian exports to LAFTA were obtained here by adding up the figures on Brazilian exports to LAFTA countries. Unfortunately, not all exports to those countries are made under LAFTA agreements.

Section 2 of chapter 5 is an interesting analysis of Brazilian export growth assuming constant market shares. The meaning of this analysis for export performance is not clear. The author calls attention to some of the limitations, especially with reference to the interpretation of the competitive effect. Under the hypothesis of constant market shares, he concludes that there was a large increase in the competitiveness of Brazilian exports from 1955–64 to 1964–71. Nevertheless, it is interesting to note also that for 1964–71 the commodity composition effect offset the gains in terms of efficiency. Does that imply that Brazil became more efficient in producing her exports but chose to export the wrong goods to the wrong market? I discuss this below.

Chapter 6 deals with the employment effect of manufactured exports, but the main question considered does not seem to be the most relevant. The author writes: "This chapter examines the possibility that Brazil might obtain substantial relief from its surplus labor problem through the promotion of its manufactured exports." He estimates that by 1994 (i.e., 23 years from 1971) manufactured exports in Brazil will generate about 50 percent of the total increase in new employment, but the manufacture exporting activities will account for about 9.6 percent (direct plus indirect) of total employment. However, the interesting aspects of the analysis of the labor content of exports and import substitutes are related to the possibility of directing policies to import substitution or export promotion trade strategies. These aspects of the results obtained in this chapter are not well explored. Some statements and conclusions are based on previous studies done by the author but not explained by him here.

Chapter 7 describes and discusses export policies and the policy environment for manufacture exports and gives a quite complete picture of the situation. Chapter 8, "The Determinants of Manufactured Export Performance," gives promise of presenting new data from the author's extensive survey work, but unfortunately this expectation is not fulfilled. The empirical results are in fact based on previously published data and consist of estimating export supplies for manufactures as a whole, not identifying the relative importance of the policy instruments discussed in chapter 7. Some qualitative inferences are based on the author's survey, which turns out to be a collection of personal interviews by the author of a few entrepreneurs, whom the author himself regards as of poor quality: "It is considered that many of these problems [general obstacles to exports] revolved around the relative scarcity of organization ability and entrepreneurial talent that seems to characterize much of the Brazilian sector."

The author to the contrary, there is reason to believe that entrepreneurs in Brazil are rather clever, since they have managed to turn the government into an important partner in sharing the costs of promoting exports.

Finally I call attention to an important technical problem. In chapter 7, and apparently throughout, the author adjusts the exchange rate by the dollar inflation to obtain the real exchange rate. This is quite misleading since the composition of Brazilian exports by destination has changed drastically against the dollar area. If we consider the average relative distribution of Brazilian exports in the period 1968–73 (United States 25.3 percent, West Germany 9.0 percent, Holland 7.0 percent, Italy 6.5 percent, Japan 5.0 percent, England 4.5 percent, France 4.0 percent, and Belgium 3.0 percent), we may infer that the index of the real rate of exchange of the cruzeiro depreciated by 9.8 percent from the first quarter of 1968 to the first quarter of 1973 (see *Conjuntura Econômica* 1975). This result contrasts with the 1.55 percent devaluation for the same period implied by the data presented in table VII-5 (p. 220) of the book. This could to some extent explain why the data in chapter 5 imply that Brazil was exporting the wrong goods to the wrong markets.

To summarize, the book may be of some interest to people that know little about Brazil. It represents a great effort to cover many aspects of a complicated subject, but its quality suffers in consequence.

José L. Carvalho

Fundação Getúlio Vargas—EPGE

References

Chenery, H. B. "Patterns of Industrial Growth." *A.E.R.* 50 (September 1960): 624–54.

Conjuntura Econômica (March 30, 1975), p. 60.

Morley, Samuel A., and Smith, Gordon W. "On the Measurement of Import Substitution." *A.E.R.* 60 (September 1970): 728–35.

Nerlove, Marc. *Estimation and Identification of Cobb-Douglas Production Functions.* Amsterdam: North-Holland, 1965.

Editors' Notice

The editors of the *Journal of Political Economy* announce a special supplement, to be published in 1979, on the work of our late colleague and fellow editor, Harry G. Johnson.

The Gibson Paradox and Historical Movements in Real Interest Rates

---

Robert J. Shiller and Jeremy J. Siegel

University of Pennsylvania

This paper analyzes the correlation between interest rates and prices which has persisted for the past quarter of a millennium and has been termed the Gibson Paradox. Spectral techniques confirm the correlation between long-term interest rates and prices for very long-term swings (the Gibson Paradox), but indicate a significant short cycle correlation only for short-term interest rates, which we term the Kitchin Phenomenon. Past explanations of these correlations have often failed to distinguish cycle lengths and term of interest rates involved. Our analysis rejects Irving Fisher's "price expectation" explanation and the Sargent–Wicksell velocity of money explanations. We propose alternative explanations which in part relate to the characteristic behavior of governments during wartime and in part to distributional effects of unanticipated inflation. Our analysis strongly suggests that prior to World War I nominal long and short rates of interest can be regarded as real rates.

For the past quarter of a millennium for which British data are available, there is a strong positive correlation between the price level, as measured by a (log) price index, and the long-term interest rate, as measured by the yield to maturity of long term bonds. This phenomenon is named by Keynes (1930) the "Gibson Paradox," after A. H. Gibson (1923), the man Keynes thought had discovered the relationship. The long-term interest rate R and the log of the price level P as we measure them are plotted together for comparison in figure 1 (see Appendix for source of data). The similarity of the two series is most impressive, especially when one

The authors are indebted to Charles Plosser for research assistance and to Anna J. Schwartz for help in assembling data. Helpful comments were made by Phillip Cagan, Michael Darby, Milton Friedman, and Franco Modigliani. Research support was provided by the National Bureau of Economic Research. Computer work was performed by the NBER TROLL system.

[*Journal of Political Economy*, 1977, vol. 85, no. 5]

© 1977 by The University of Chicago. All rights reserved.

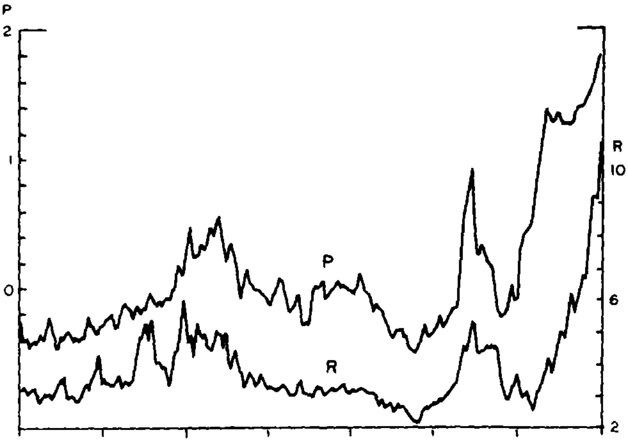


FIG. 1 The log of the price index (P) and the long-term interest rate (R), British data. See Appendix for source of data.

considers the major changes in social, political, and economic structure which took place over this period.[1] Keynes called this correlation "one of the most completely established empirical facts in the whole field of quantitative economics" (1930, 2: 198).

Over the shorter time period for which good British data on short-term interest rates are available, there is also a substantial but less pronounced positive correlation between short-term interest rates and prices. The short-term interest rate I is plotted in figure 2, with the same price series P. The correlation is not only less pronounced than that in figure 1;[2] it

---

[1] The Gibson Paradox appears in data from other countries as well. Fase (1972) concluded that the paradox holds in the Dutch economy. The paradox also appears over the shorter time period for which data are available in the United States and Canada. For the U.S. data, the relation between a long-term bond yield series and prices is again a long-term one, and shorter-run movements in the series show little relationship. One may not, of course, regard the experience of these other countries as independent observations, since their interest rates and prices are related through financial markets. There is, however, substantial variation between U.S. and British price series and U.S. and British bond yield time series. The data show that U.S. federal government bond yields were dramatically above the British consol yields during the War of 1812 and the Civil War, and these periods corresponded to periods of very high prices in the United States relative to Britain. Experiences of some other countries, however, may be construed as disconfirming the Gibson Paradox. For instance, the price level in France rose during the Second World War to a plateau which was, in the 1950s, 2,500 percent higher than in 1937. Interest rates did not rise to a new higher plateau. We could find similar counterexamples in the experience of certain Latin American countries with dramatically unstable price levels.

[2] The correlation between the R and P data in figure 1 is .743, while that between R and ΔP (the rate of inflation) is .045. The correlation between I and P in figure 2 is .421, between I and ΔP is $-.01$.

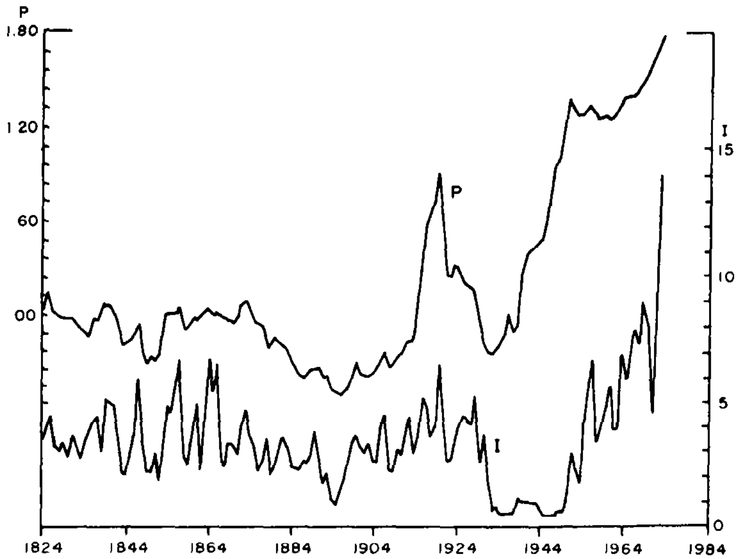


FIG. 2.—The log of the price level (*P*) and the short (3-month) interest rate (*I*), British data. See Appendix for source of data.

also takes a different form. The relation between *I* and *P* is much stronger over the short term than that between *R* and *P*. We will refer to this short-term correlation with the short rate of interest as the Kitchin Phenomenon, after Joseph Kitchin who noted it in 1923.

A correlation between interest rates and prices of this significance must certainly have been noted many times in the past, and indeed it has been an empirical regularity that has been of great importance in the development of monetary theory. The correlation was apparently first described by Tooke (1844). He used it to criticize the Thornton-Ricardo view that an increase in the money supply is transmitted to an increase in prices by a decline in the interest rate and to defend the real bills doctrine. Alternative veiws of the phenomenon were developed by Wicksell (1898, 1962) and by Hawtrey (1913). Renewed interest in the phenomenon was then prompted by the "discovery" of the relation of long rates to prices by Gibson (1923), and of the relation of short rates to prices by Kitchin (1923) and Peake (1928). Kitchin compared the peaks and troughs of *I* and *P* and noted that the turning points which occur together occur about 40 months apart, a phenomenon that has come to be known as the "Kitchin Cycle." Unfortunately, what is best known today is Kitchin's conclusion that a 40-month cycle is discernible from the *I* or *P* series individually.

A number of theories have been proposed to explain the positive correlation between interest rates and prices, notably by Fisher (1930), Keynes (1930), MacCaulay (1938), and Sargent (1973). In this paper we analyze these theories and propose alternative explanations of the Gibson Paradox and Kitchin Phenomenon.

Analysis of the British Data

Spectral analysis of the time series appearing in figures 1 and 2 gives us a more precise description of the pheomena described above:[3]

1. The squared coherence (which is essentially a squared correlation coefficient at each frequency) between R and P, as shown in the top of figure 3, is stronger at the lower frequencies, where it is significant at the 5 percent level. The squared coherence between I and P shown in the bottom of figure 3 is concentrated, in contrast, more in the higher frequencies and peaks very near the Kitchin frequency of .300 (40 months), where it is also significant at the 5 percent level. In both cases the phase angle is nearly zero wherever coherence is strong (i.e., the relations are essentially unlagged).

2. Considered individually, none of the series shows a spectrum which is very concentrated at any particular (nonzero) frequency; that is, there is little evidence of predominant business cycle for prices and interest rates. The spectra of both R and P are typical of the kind of spectra which, as elaborated by Granger (1966), are commonly found among economic time series, in that their most striking feature is the considerable power at the lower frequencies. Box-Jenkins analysis suggests that both series might be represented as integrated processes. Since Klein (1975) found major changes in the behavior of prices after the end of the gold standard, we tried dividing the sample period rather arbitrarily at 1913. As a first

---

[3] The sample period is 1730–1973 in all estimates except those involving I, for which a sample period of 1826–1937 was used. The shorter sample period was necessitated by lack of good data before 1820 and by the unusual controls on short rates around World War II, whose effect in drastically lowering short rates can be readily seen in figure 2. In computing the cross-spectra, the data were quasi first differenced (with filter [1 − .8L]), and then Fourier transformed (without prior padding with zeroes). The cross-periodogram was smoothed with a wrap triangular filter. In computing the cross-spectra between R and P the data were initially detrended, but this was unnecessary for I and P. The range of the filter for R and P was 18, and the band width was .065. The range of the filter for I and P was 12, and the band width was .102. Since Box-Jenkins analysis suggested that R and P may be represented as random walks for the greater part of the sample period, the series may not be stationary, so that their spectra would not be defined. Even if this is so, the spectra remain useful characterizations of the data. Using the filter (1 − .8L) on all the data produced spectra of R and P which were whitened considerably (and a spectrum of I with much reduced lower frequencies), so that the coherence pattern of the first differenced series would not be substantially different.

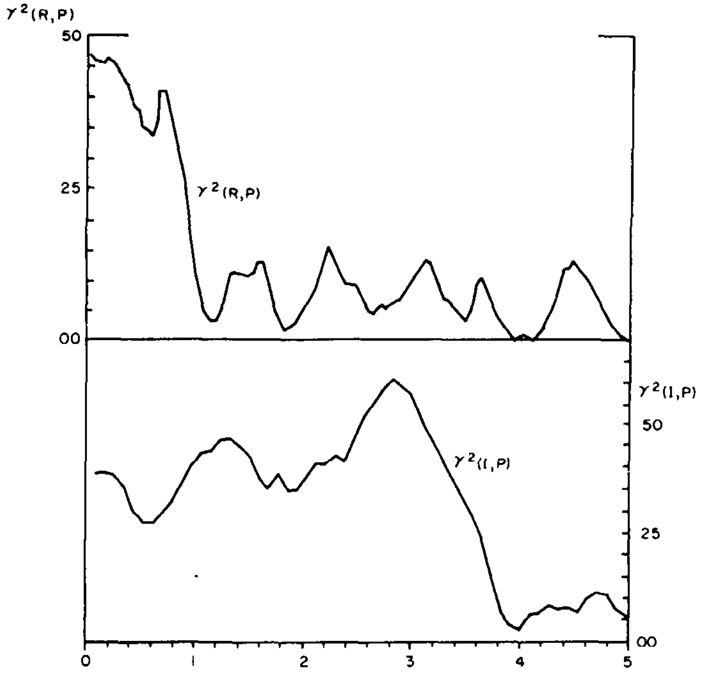


FIG. 3. Plots of squared coherence (vertical axis) against frequency in cycles per year (horizontal axis). Upper plot: squared coherence between *R* and *P*. Critical coherence squared at 5 percent level is .350. Lower plot: squared coherence between *I* and *P*. Critical coherence squared is .472. See n. 3 for details of estimation.

approximation, it is useful to regard both of these series as random walks for the period up through 1913, and as positively serially correlated IMA (1, 1) (integrated moving average processes) for 1914–74. Thus, as Klein emphasized, our impression that inflation is forecastable is valid for the present time, but not for most of the period covered by our data. The appearance to the eye of "trends" in figure 1 before 1914 is the usual illusion given by random walks. These Box-Jenkins representations, however, do not perfectly correspond to the observed spectra. The long rate spectrum also shows small humps in the middle frequencies, peaking at frequencies .23 and .32. The short rate series *I*, in contrast to the above, has a spectrum much less concentrated in the low frequencies, as evidenced by the relatively choppy appearance of the series in figure 2. Again, there are no important humps in the spectrum. There is a small hump in the Kitchin frequency of .30, but other humps which are just as large occur at lower frequencies.

Theories Based on a Response in Interest Rates to Inflationary Expectations

Irving Fisher in his classic *The Theory of Interest* (published in 1930) has given the most popular explanation of the Gibson Paradox. Fisher stated that the market rate of interest at time t, R_t, on long-term bonds is the sum of a more or less constant real rate of interest $\bar{\rho}$ plus an expected "long run" rate of inflation at time t, π_t, so that $R_t = \bar{\rho} + \pi_t$. He asserted that the Gibson phenomenon arises because π_t is positively correlated with the price level, and he attempted to provide a plausible theory of expectations which would give this result.[4]

According to this theory, if we assume knowledge of future prices the price V_t of a consol paying coupons of C dollars each time period is the present value of its real coupons (in time t prices) discounted by a constant real rate $\bar{\rho}$. The yield $R_t = \log(1 + C/V_t)$ of the consol can be approximated by linearizing around $\Delta P(t + n) = 0$, $n = 1, 2, \ldots$ (see Shiller 1972) to obtain

$$R_t = \bar{\rho} + [1 - \exp(-\bar{\rho})] \sum_{k=1}^{\infty} \exp(-k\bar{\rho}) \Delta P_{t+k}. \tag{1}$$

Thus, consol yields are given approximately as $\bar{\rho}$ plus a weighted average of future rates of inflation, the weights declining geometrically with time. This weighted average we refer to as an "expected long-run rate of inflation." In effect, the long-run rate of inflation is related to the "present value" of future one-period rates of inflation discounted by the real rate of interest.

Fisher claimed that expected long-run inflation rates π_t could be represented as a distributed lag on past actual inflation rates. If the distributed lag is exponential, this gives

$$R_t = \bar{\rho} + b \sum_{k=1}^{\infty} \exp(-ak) \Delta P_{t-k}, \tag{2}$$

a model which resembles (1) except that the weighted average includes lagged ΔP rather than led ΔP. As an expectations mechanism, it sounds superficially plausible that individuals should form expectations as a weighted average of lagged inflation rates. Moreover, if a is very small we can be sure that the model will fit the data very well since

$$\sum_{k=1}^{\infty} \exp(-ak) \Delta P_{t-k}$$

---

[4] The Fisher theory has spawned an enormous literature (see, for example, Roll 1972, Modigliani and Shiller 1973, Sargent 1973, and Fama 1975).

is essentially the price level itself, thus such a "model" becomes a simple restatement of the Paradox.[5] We know that R will fit π so defined very well.[6]

Although the model fits well, however, it is not a plausible one from the standpoint of investor rationality. To see this, we look at expression (1). Since R here is a linear function of P_t, $P_{t+1}, \ldots$ we can derive the cross-spectrum between R and P.[7] It turns out that the phase angle from P to R is for high frequencies (short cycles) near 180°, implying that high-frequency components of P are matched by high-frequency changes in R in the opposite direction. The intuitive explanation is simple: in the short-run the price of the bond must move with the price level for real returns to be kept constant for short holding periods. In short periods of time the coupon payments are negligible and can be disregarded. Keynes suggested this point in his *Treatise on Money*. For long cycles, on the other hand, the phase angle approaches 90°. At these frequencies yield becomes uncorrelated with the price level. Hence, there is thus nothing to suggest a phenomenon like the Gibson Paradox.

A more graphic way to see the "paradox" involved here is to compute what interest rates would have been had investors known with certainty what future prices actually transpired. We will define an "ex post rational yield" as the yield on a bond whose price is the present value of its real coupons discounted by $\bar{\rho}$. If R^* is the yield on such a consol, then we have from the definition of yield the recursive relation

$$R_t^* = \log \{1 + \exp(\bar{\rho} + \Delta P_{t+1}) \cdot [1 - \exp(-R_{t+1}^*)]\}. \qquad (3)$$

Given a value for $\bar{\rho}$ and a terminal value for R^*, which we took to be 3 percent and the actual yield for 1973, respectively, we can compute an ex post rational yield for all earlier dates for which we have price data. This is plotted at the top of figure 4. This ex post rational yield guarantees that holding period real returns for *all* periods be 3 percent. Clearly the ex post rational yield moves *opposite* actual yields, which appear in the middle of figure 4. The correlation between ex post rational yields and actual yields from 1729 to 1950 (a period chosen to exclude the most

---

[5] For small a the theoretical gain from P to R according to (2) is essentially flat and the phase angle zero for all but the lowest frequencies.

[6] There is an extensive literature on the estimation of the parameters of (2), although the relation is often modified by making $\bar{\rho}$ a function of other variables (see Roll 1972). Some authors have not found the distributed lag significant for certain subperiods and estimation procedures (see Cargill and Meyer 1974).

[7] The phase angle is given by

$$\phi(w) = \arctan \left\{ \frac{(1 - e^{-\bar{\rho}})}{(1 + e^{-\bar{\rho}})} \cdot \frac{\sin(2\pi w)}{[\cos(2\pi w) - 1]} \right\}$$

and the gain is given by

$$g(w) = (1 - \text{[illegible]}^{-\bar{\rho}})\{[2 - 2\cos(2\pi w)]/[1 - 2e^{-\bar{\rho}}\cos(2\pi w) + e^{-2\bar{\rho}}]\}^{.5}.$$

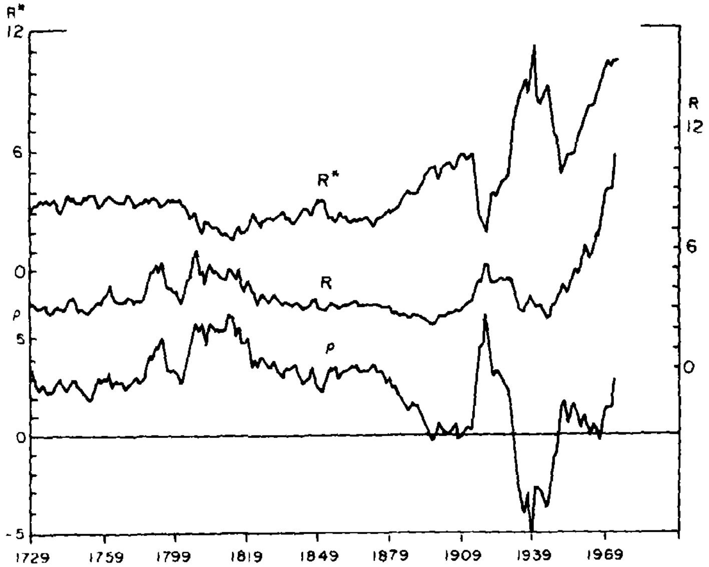


FIG. 4.—Ex post rational yields R^*, actual yields R and ex post real long-term rates of interest ρ. British data.

recent values of our ex post rational yield which are heavily influenced by the arbitrary terminal value chosen) was −.30.

The ex post *real* yield on a bond is that rate of interest which discounts the real coupons to the *actual* real value of the bond. In the case of a consol whose price is V_t and which carries coupon C per time period, the ex post real yield ρ_t at time t is defined implicitly by

$$V_t = C \sum_{n=1}^{\infty} \exp\left(-\rho_t n - \sum_{m=1}^{n} \Delta P_{t+m}\right). \tag{4}$$

This implicit function for ρ_t in terms of C, V_t, and future ΔP cannot in general be solved for ρ_t. However, if inflation rates are not large, then, based on a linearization around $\bar{\rho}$, we derive the approximation

$$\rho_t \simeq R_t - (R_t^* - \bar{\rho}) \tag{5}$$

where $\bar{\rho}$ is the real rate used in the computation of R^*, the ex post rational yield. The ex post real long rate is here essentially the nominal rate minus the ex post long-run rate of inflation which is implicit in R^*. The values of ρ_t computed from the R^* at the top of figure 4 are plotted at the bottom of figure 4. By construction, the real rate approaches $\bar{\rho}$ (3 percent) at the final date. The standard deviation from 1729 to 1950 of ρ is 217 basis points. If R_t had been constant over this period, then using R^* and (5) we compute that the standard deviation of ρ over this period would have been 182 basis points; that is, as we have noted, the movements in R aggravated, rather than mitigated, the effect of changes in P on real yields.

The ex post real long rate of interest is strongly negative in the depression of the 1930s which means that the real value of the bond then was greater than the simple sum of its real coupon payments up to 1973 plus its real value in 1973. Real rates of interest hit zero during the depression of the 1890s, and were relatively low during the last part of the nineteenth century. The secular deflation then was not severe enough to offset the low nominal yields.

Of course, individuals do not possess perfect foresight. Based on the stochastic properties of the price level alone, before 1914 investors should have projected a near zero rate of inflation consistently, since the price level displayed little upon which to project future rates of inflation. When there were external circumstances that should have given rise to nonzero projections, as during the war years 1810–15, the Napoleonic Wars, and especially the 1920 World War I peak, it would appear that investors consistently forecast prices in the wrong direction. The same could be said for the depression period of the 1930s and the governmental suppression of inflation during World War II.

In conclusion, the Fisherian hypothesis that price expectations account for the Gibson Paradox and the real rate is constant cannot be maintained on the basis of a rational expectation mechanism for generating inflationary expectations. If the price level follows a random walk and expectations of future inflation are based only on past inflation, then the interest rate series we see is in fact the real rate of interest.

Theories Which Rely on Interest Sensitive Demand or Supply of Money

A completely different approach to the explanation of the Gibson Paradox was taken by several other authors, who combined a characterization of an interest-sensitive demand for supply of money with a characterization of the time path of exogenous variables. Keynes (1930) and Wicksell (1962) proposed explanations which, while differing in details, both relied on an interest-sensitive money supply. Sargent (1973) proposed a model of the Gibson Paradox which relied instead on an interest-elastic *demand* for money. We may summarize both types of explanation in terms of the two-equation model

$$\log(M) = \log[\mu(r)] + \log(H), \qquad \text{money supply} \qquad (6)$$

$$\log(M) + \log[V(r)] = P + \log(y), \qquad \text{money demand} \qquad (7)$$

where M is the money supply, $\mu(r)$ is the interest-sensitive money multiplier, $\mu'(r) > 0$, H is the supply of high-powered money, $V(r)$ is the interest-sensitive velocity of money, $V'(r) > 0$, and y is real national product. The interest rate r could be either the short-term interest rate I or the long-term interest rate R. Let us for the moment assume it is R. If we assume demand equals supply we may derive from (6) and (7) the reduced form for P

$$P = \log [V(r)] + \log [\mu(r)] + \log (H/y)$$

or (8)

$$P = \log V_H(r) + \log (H/y)$$

where $\log V_H(r) = \log [V(r)] + \log [\mu(r)]$ is the velocity of high-powered money. This result divides the log of the price level into two or three components, all components are directly observable, $V(r) = Y/M$ where Y is money income, and $[\mu(r)]$ is M/H. Sargent's model asserts that $\log [\mu(r)]$ and $\log (H/y)$ are relatively constant, and the Gibson Paradox arises because $V(r)$ moves with R. Wicksell's and Keynes's models assert that $\log V(r)$ and $\log (H/y)$ are relatively constant, and the Gibson Paradox arises because $[\mu(r)]$ moves with R. All of these theories assert that the variation in P must be due primarily to variations in the velocity of high-powered money $V_H(r)$. One way of evaluating these theories is thus to check which of the components of P in (8) account for movements in P.

Before we do this, we must decide on the definition of high-powered money H. If we take the conventional definition which includes Bank of England notes as well as gold, then it is neither $V(r)$ nor $\mu(r)$ which does most to explain P; it is instead the third component, $\log (H/y)$, as can be seen from figure 5. The correlation between P and $\log (H/y)$ is an extraordinary .972. The log of the velocity of high-powered money $\log (Y/H)$ plotted at the bottom of figure 5 accounts for only a small part of the variation in P. This is not to say that the velocity of high-powered money is not sensitive to the interest rate, since the correlation coefficient between the consol yield series and the log of Y/H is .348, but the changes in this velocity are not very important in explaining price movements compared to changes in the money stock. In particular, one notes that the high prices and interest rates during World War I cannot be associated with an increase in the velocity of high-powered money.[8]

The data thus indicate that neither Sargent's theory nor Wicksell's theory can be the principal explanation of the Gibson phenomenon. On the other hand, if we define high-powered money to be reserves of the Bank of England, as Keynes at least clearly intended in this context, then these data are entirely consistent with Keynes's explanation. The possibility that Keynes suggested is that governments or central banks will

---

[8] Cagan (1965) has also emphasized the importance of H/y in determining P based on U.S. data.

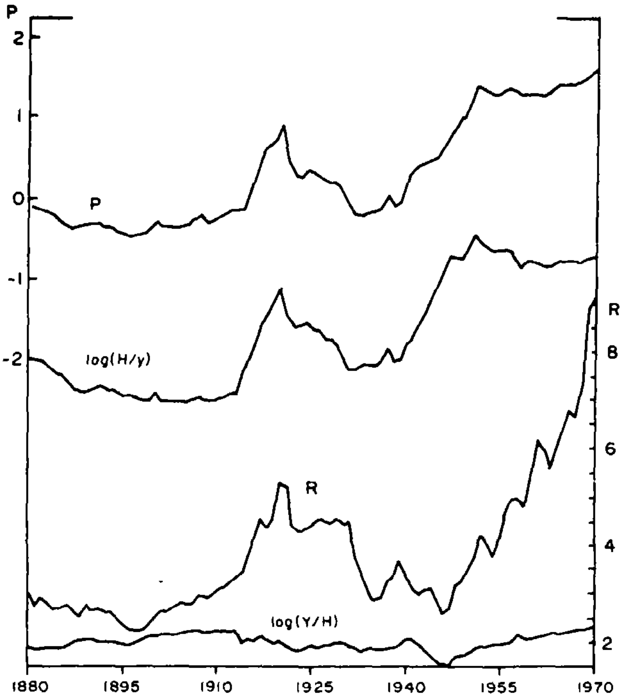


FIG. 5.—Upper plots: the log price level P and the log of the ratio of high-powered money to real GDP, log (H/y). Scale for these plots is at left. Lower plots: The long-term interest rate R and the log of the velocity of high-powered money, log (Y/H). Scale for these plots is at the right.

respond to increases in interest rates by increasing the money supply.[9] If the central bank seeks, with less than complete success, to stabilize the long-term interest with its handling of the money supply, then the Gibson phenomenon might emerge. Keynes (1930) argued, as an element of his explanation of the Gibson Paradox, that central banks have long behaved in this manner.

The clearest case for such behavior occurs during wartime. Debt-financed war expenditures are an attempt to shift part of the burden of the war to the future. To the extent that individuals do not discount all

[9] In the United States also, even before the Federal Reserve System, the government has had influence over high-powered money. In the Civil War, the U.S. printed "greenbacks" not immediately redeemable in gold. Moreover, during the War of 1812 (even though there was no central bank, the charter of the Bank of the United States having expired in 1811) banks in the middle states had suspended specie payments and the federal government was able to obtain loans of their unredeemable paper notes from them, thus effectively increasing the money supply.

future tax liabilities arising from the issuance of debt, current government borrowing will not be met by an equivalent increase in saving unless the real rate rises to equilibrate the financial market. The central bank may succumb to the political pressure of keeping the interest rate from rising by increasing high-powered money and hence causing inflation. Thus, during wartime the correlation between interest and prices arises because each, in turn, is causally linked to two other variables that are correlated: government interest-bearing debt and high-powered money.[10]

This explanation is capable of explaining the Kitchin Phenomenon as well as the Gibson Paradox. Since the coherence between P and the log of H/y is high for all frequencies, both the low-frequency coherence between P and R and the high-frequency coherence between P and I could be explained with an appropriate model linking H/y to both I and R. If we suppose that the monetary authority attempts to stabilize short-term cycles in the short rate of interest and long-term cycles in the long rate of interest, then both the Gibson and Kitchin phenomena would emerge. This type of policy might arise if the central bank responds only to large movements in either series, since our spectral analysis demonstrates that large movements in R are associated with long cycles and large movements in I with short cycles.

Theories Which Depend on Effects of Unanticipated Price Changes on Contracts Made in Nominal Terms

Changes in the price level which do not affect relative prices should have no real effects and should hence not affect the real interest rate if no intertemporal contracts are made in nominal terms. Traditionally, a very large variety of contracts are made specifying payments to be made at different times which are denominated in money terms. When such contracts are made, unanticipated price changes have the effect of redistributing income across individuals and possibly inducing real effects.

We now develop a model in which these price-induced wealth redistributions can give rise to the Gibson Paradox or Kitchin Phenomenon.[11]

[10] Hawtrey (1932) observed this tendency clearly: "Governments do not want to borrow from their central banks (apart from quite legitimate overdrafts for a few days) except at a time of overwhelming financial strain, and at such a time the central banks *never* resist their demands. . . . In face of the exigencies of war, no reliance can be placed on an independent central bank to resist demands from the government. At such a time the government is accepted by public opinion as the sole authoritative exponent of the national needs. If the directors of the central bank raise objections to the government's financial measures, they will be overborne."

[11] Much of this explanation is drawn from Siegel (1975). Other price-induced distributional explanations have relied on the existence of fixed nominal wages. Hawtrey (1913) theorized that unanticipated price changes result in shifts of wealth between workers and entrepreneurs because of sticky money wages in such a way as to cause corresponding shifts in capital productivity and the return on capital, but his theory

In this model, we assume that the economy is initially in stochastic equilibrium. Individuals will be divided into two groups. One group (demanders) wishes to hold net long positions in nominally denominated assets such as consols. For simplicity we shall assume that the net supply of these assets is zero, so a second group (suppliers) must necessarily be short in these assets. For instance, a firm whose real capital stock is partially financed by bonds has the property that equity holders are net short in those bonds. Assume that the interest rate on bonds equilibrates the demand and supply.

When the price level has risen unexpectedly, the wealth position of those long in bonds will have decreased, and the wealth position of those short will have increased by the same amount. Assume that each group wishes to maintain a certain fraction of its total wealth in long or short positions in bonds. An unexpected price rise will reduce the real supply of nominal debt in existence, but the demanders of nominal debt will seek to increase their bond positions by only a fraction of this change in supply once their wealth has decreased. The net suppliers of debt will wish to increase the supply by more than the drop in the value of the bonds since their wealth position has been enhanced. Hence, at the initial equilibrium interest rate there will be an excess supply of bonds, and the interest rate will have to rise to equilibrate the market.

We may, although it is not necessary, associate those who demand bonds as relatively "risk averse" and those who supply bonds (lever their investments) as relatively risk taking. If this is the case then unexpected changes in the price level change the relative wealth positions of the relative risk averse and relative risk takers in the economy. An unexpected upward movement in prices would tend to change the risk characteristics in the economy toward the preferences of the risk takers and conversely in the case of unexpected price declines.

A change in the rate of interest on bonds hence may not be associated with any real change in the productive technique or quantity of capital in the economy,[12] but only reflect changes in the income distribution the economy caused by unanticipated price movements. An increase in, say, risk aversion resulting from an unanticipated price fall will cause "sure" claims to trade at a higher price (and lower rate) than risky capital. This last point is particularly important since, if we interpret the market rate of interest as the real rate of interest (as the data up to World War I suggest) then a deterministic model will suggest results that may not exist

---

appears poorly developed. MacCaulay (1938), who resolutely rejected Irving Fisher's explanation, felt that distribution effects, through their influence on levered firms' profitability, were responsible for the Gibson Paradox. It should be noted, as Keynes (1930) so aptly states, that any short-run or business cycle explanation of the Gibson Paradox is insufficient due to its long-run nature.

[12] *For a more detailed discussion of this, see* Siegel *and* Warner (1977).

in a stochastic world. A deterministic model would predict that the capital-output ratio in the economy would move inversely to the real rate of interest, and hence we would expect that the capital-output ratio would conform inversely to the historical interest rate pattern. Available evidence does not suggest any such countercyclical movements of capital intensity—no deepening during the latter half of the nineteenth century or "thinning" thereafter. Certainly during the depression of the 1930s the capital stock did not rise. As mentioned above, changes in aggregate distribution of income (and possibly risk preferences) can bring about such changes without any concomitant changes in the capital stock.

A result of this theory is that the rate of return on real, unlevered equity need not follow the pattern of the long-term nominal bond rate even if this rate corresponds to the real rate of interest. Interestingly, evidence that supports constant unlevered returns also supports one of Fisher's assumptions: that expected real rates were constant. We are also suggesting in this analysis that there may be a constant expected real asset return, but differ from Fisher in that we suppose that the expected real risk-free rate is one which will fluctuate with the (unanticipated) movements of the price level due to distribution effects and not inflationary anticipations.

Our distribution-effects explanation is consistent with the coherence pattern which characterizes the Kitchin Phenomenon, if we add the assumption that there is some, say, white noise affecting supply and demand for credit and hence I, the short-term interest rate. The distributional effects of unanticipated price changes are likely to be temporary, if for no more reason than that generations are finite lived. The high power of P at the lowest frequencies is thus not transmitted to I, and so because of the other noise in I at those frequencies coherence is low there. Moreover, since the power of P is low at the highest frequencies, it is also probable that the hypothesized noise in I might swamp out the influence of P on I at these frequencies, again producing low coherence. Thus, the inverted U-shaped coherence pattern at the bottom of figure 3 which is high only at the middle frequencies might well emerge.

We might also assume some noise component in the supply and demand for long-term bonds to explain why coherence is low for the higher frequencies between R, the long-term bond rate, and P. It is more difficult, however, to make a case for the concentration in the extreme low frequencies of the coherence between R and P. To explain this, we would have to assume that the effect of movements in P on the distribution of wealth is attenuated very slowly.

Conclusion

Our analysis confirms a long-cycle correlation between *long-term* interest rates and prices, popularly termed the Gibson Paradox, and a short-cycle correlation between *short-term* interest rates and prices which we term the Kitchin Phenomenon. Long-term interest rates and prices display no significant short-cycle correlation; hence any "business-cycle" explanations of the Gibson Paradox are invalid.

Our results suggest that it is probably unreasonable to suppose, as many have, that historical changes in long-term interest rates can be attributed largely to changes in inflationary expectations. Thus, historical movements in nominal yields are probably a reflection of movements in the expected real long-term rate of interest. Second, it is also unreasonable to attribute historical movements in the price level to changes in the velocity of money. Long-run movements in the price level arise instead primarily from changes in the supply of high-power money, so an explanation of the Gibson Paradox which relies on interest rate induced shifts in velocity or the money multiplier cannot be maintained. However, it is possible that central bank behavior, in attempting to attenuate interest rate changes, will result in a correlation between high-powered money and interest rates and hence give rise to the Gibson Paradox.

Another possible explanation involves changes in interest rates that arise due to the effects of unanticipated changes in the price level on contracts made in money terms. A price rise distributes wealth from creditors to debtors, which changes the interest rate which equilibrates demand and supply for nominally denominated assets. The extent to which this explanation explains the Gibson Paradox as well as the Kitchin Phenomenon depends on how quickly economic agents readjust altered wealth positions.

Appendix

Sources of Data

Price data come primarily from Mitchell and Deane (1962) and Mitchell and Jones (1971). The Wholesale Price Index (whose log is P) for 1846–1966 is the Sauerbeck-Statist Overall Price Index, which is the index used by Keynes in his discussion of the Gibson Paradox. The series was extended by linking to it other series, multiplied by the ratio of the two series for 1 overlapping year. In this manner, the Gayer Rostow and Schwartz Monthly Index of British Commodity Prices was linked to it for the years 1790–1846. To this in turn was linked Elizabeth Schumpeter's Consumer Goods Price Index for the period 1727–90. Since the Statist series was discontinued in 1966, the wholesale price index number of output of manufactured products, total sales from the Central Statistical Office, was linked to the Statist Index for the period 1967–73.

The yield series R is basically that on British Perpetual Annuities from 1727 to 1973, compiled by Homer (1963, tables 13, 19, and 57). Beginning with 1753, there are yields of 3 percent consols, redeemable at the option of the government. In the 1880s yields fell below 3 percent, so that the yields were prevented from falling much further by the possibility of redemption. After the refunding of 1888, the 3 percent consols were replaced with new consols yielding 2¾ percent until 1903, and thereafter 2½ percent. Since the yields on consols were thus abnormally high in the 1880s, we here follow Homer's suggestion and use yields on 2½ percent government annuities for the years 1881–88, consol yields thereafter.

Bill rates I are taken primarily from Mitchell and Deane (1962, p. 460). From 1824 to 1844 we use Overend and Gurney's rates from Mitchell and Deane (1962) for first-class 3-month bills. From 1845 to 1938 we use the average of the annual means of the monthly means of the highest and lowest rates of each day on 3-month bank bills (Mitchell and Deane 1962, p. 460, part B). From 1939 to 1965 we use the annual mean of the market buying rate on working days of 3-month bank bills quoted in Mitchell and Jones (1971, p. 182). The most recent data were the 3-month prime bank bill rate averaged over the year, from the Central Statistical Office Abstract of Financial Statistics.

Gross Domestic Product (Y) for the United Kingdom for the years 1870–1965 are those given in Feinstein (1972, table 3, col. 9, pp. T10–T11). Real Gross Domestic Product (y) was formed from Y by dividing Y by our price index P.

The high-powered money supply series H and the money supply series M are basically those given in Sheppard (1971). The actual data we used, however, were supplied to us by Anna J. Schwartz. Her series involves some minor judgmental changes and updates Sheppard's series based in part on recent data supplied her by Sheppard.

References

Cagan, Phillip. *Determinants and Effects of Changes in the Stock of Money 1875–1960.* New York: Columbia Univ. Press (for Nat. Bur. Econ. Res.), 1965.

Cargill, Thomas F., and Meyer, Robert A. "Interest Rates and Prices since 1950." *Internat. Econ. Rev.* 15 (June 1974): 458–71.

Fama, Eugene F. "Short Term Interest Rates as Predictors of Inflation." *A.E.R.* 65 (June 1975): 269–82.

Fase, M. M. G. "Bond Yields and Expected Inflation." *Econ. Q. Rev.* 30 (September 1972): 5–10.

Feinstein, C. H. *National Income Expenditure and Output of the United Kingdom.* Cambridge: Cambridge Univ. Press, 1972.

Fisher, Irving. *The Theory of Interest.* New York: Macmillan, 1930.

Gibson, A. H. "The Future Course of High Class Investment Values." *Banker's Magazine* (London) 115 (January 1923): 15–34.

Granger, C. W. J. "The Typical Spectral Shape of an Economic Variable." *Econometrica* 34 (January 1966): 150–61.

Hawtrey, R. G. *Good and Bad Trade.* London: Constable, 1913.

———. *The Art of Central Banking.* London: Longmans, Green, 1932.

Homer, Sidney. *A History of Interest Rates.* New Brunswick, N.J.: Rutgers Univ Press, 1963.

Keynes, J. M. *A Treatise on Money.* New York: Macmillan, 1930.

Kitchin, Joseph. "Cycles and Trends in Economic Factors." *Rev. Econ. Statis.* 5 (January 1923): 10–16.

Klein, Benjamin. "Our New Monetary Standard: The Measurement and Effects of Price Uncertainty, 1880–1973." *Econ. Inquiry* 13 (December 1975): 461–83.

MacCaulay, Frederick. *Some Theoretical Problems Suggested by the Movements of Interest Rates, Bond Yields, and Stock Prices in the United States since 1856.* New York: Nat. Bur. Econ. Res., 1938.

Mitchell, B. R., and Deane, P. *Abstract of British Historical Statistics.* Cambridge: Cambridge Univ. Press, 1962.

Mitchell, B. R., and Jones, H. *Second Abstract of British Historical Statistics.* Cambridge: Cambridge Univ. Press, 1971.

Modigliani, Franco, and Shiller, Robert. "Inflation, Rational Expectations, and the Term Structure of Interest Rates." *Economica* 40 (February 1973): 12–43.

Peake, E. G. *Bankers Magazine* (London) 125 (May 1928): 720.

Roll, Richard. "Interest Rates on Monetary Assets and Commodity Price Index Changes." *J. Finance* 27 (May 1972): 251–77.

Sargent, Thomas J. "Interest Rates and Prices in the Long Run: A Study of the Gibson Paradox." *J. Money, Credit and Banking* 4 (February 1973): 385–449.

Sheppard, David K. *The Growth and Role of UK Financial Institutions.* London: Methuen, 1971.

Shiller, Robert J. "Rational Expectations and the Structure of Interest Rates." Ph.D. dissertation, Massachusetts Inst. Tech., 1972.

Siegel, Jeremy J. "The Correlation between Interest and Prices: Explanations of the Gibson Paradox." Unpublished manuscript, Univ. Pennsylvania, 1975.

Siegel, Jeremy J., and Warner, Jerold. "Indexation, the Risk-Free Asset, and Capital Market Equilibrium." *J. Finance* (September 1977).

Tooke, Thomas. *An Inquiry into the Currency Question.* London: Longman, 1844.

Wicksell, Knut. *Geldzins and Guterpreise* [*Interest rates and prices*]. Jena: Fischer, 1898.

———. *Lectures on Political Economy.* London: Routledge & Kegan Paul, 1962. (Lectures given in 1911.)

The Relevance of the Two-Sector Production Model in Trade Theory

---

Ronald W. Jones
University of Rochester

José A. Scheinkman
University of Chicago

This paper examines how well the basic properties of the traditional 2 × 2 model of a competitive economy, commonly used in much of the pure theory of international trade, generalize when more goods and factors are considered. The notion of factor intensity and the Heckscher-Ohlin, Stolper-Samuelson, and Rybczynski theorems are discussed. The role played by the no-joint-production assumption as opposed to small dimensionality in the latter two results is stressed. The mathematical appendix provides a compact and formal statement of the properties discussed in the text.

The two-sector, two-factor model of a competitive economy is by now standard fare in a number of fields in economics. In particular, the pure theory of international trade has made extensive use of the "2 × 2" model. As in any body of theory, the results and theorems obtained rest upon the entire set of assumptions. How crucial is the assumption limiting the number of factors and commodities to two apiece? The widespread use of the model suggests that the prevailing belief is that basic properties and theorems in some sense are capable of generalization. On the other hand, there are those who are strongly critical of the simple model precisely because its dimensionality is limited to the 2 × 2 case.

Consider the following sample of critical opinion. We begin with Ivor

We wish to thank Ikushi Egawa, Winston Chang, Trout Rader, and Harry Johnson for comments on this paper as well as Bill Ethier and Murray Kemp for earlier remarks in 1973 on preliminary work. A recent paper by Diewert and Woodland (1977) provides an elegant model with some results in common with our own paper. This research was supported by the National Science Foundation.

[*Journal of Political Economy*, 1977, vol. 85, no. 5]
© 1977 by The University of Chicago. All rights reserved.

Pearce (1970), who devotes much of his book to challenging the usefulness and general validity of the simple traditional trade models. Thus ". . . we shall again and again argue that it is in fact not possible to study problems involving factor prices *in anything like a realistic way* until at least three commodities and three factors are introduced," or, further on, ". . . many textbooks of international trade theory (even the most advanced) lay a great deal too much emphasis upon ***propositions which are true only for models with two commodities and two factors of production***" (p. 320, italics supplied). The latter quote is quite explicit in saying that at least some of the standard propositions are false for economies with more than two commodities and factors. This of course goes further than the remark that it may be difficult to prove results in the case of many goods and factors without adding further assumptions. Murray Kemp (1971), in a review of Pearce's book, states with approval, "He shows that many of the familiar propositions of two-by-two-by-two trade theory do not carry over to larger and, from a policy-making point of view, more relevant systems" (p. 251). Finally, consider the following remark by Frank Hahn (1973): "It is well known that an economy with only two goods has a number of important properties which do not carry over to the general case" (p. 297).

The purpose of this paper is to respond to these charges by considering how well or poorly the basic properties of the 2 × 2 model generalize. In the first section of the paper, we present the basic set of four propositions in the pure theory of international trade as typically set out for the standard model. This is followed by a section discussing a generalized notion of factor intensity, a concept central to the Heckscher-Ohlin theorem about trade flows. Section III establishes a general form of the dual Stolper-Samuelson and Rybczynski theorems and points out that the crucial assumption upon which these properties rest is *not* small dimensionality but, rather, the lack of joint production.

Two features concerning dimensionality are found in the 2 × 2 model. The first is that 2 is a small number. The second is that the number of factors is the same as the number of commodities. Section IV of the paper discusses those properties of a general model that hold if the number of factors and commodities are the same (the "even" case) but may not hold if there are more factors than commodities (the "uneven" case). The concluding section discusses some basic properties found not only in the 2 × 2 case but in much more general settings as well—in both "even" and "uneven" cases and in cases in which joint production appears.

The discussion in the body of the paper is kept deliberately nonmathematical, primarily to establish that the properties being discussed rest upon simple assumptions about technology. The Appendix reconsiders some of this material in a more compact matrix format.

I. The Basic Trade Model

The standard Heckscher-Ohlin model of international trade incorporates a set of propositions that reveal essential properties of the two-commodity, two-factor general equilibrium model of production. Of these propositions, four are central:[1]

i) The *Heckscher-Ohlin theorem* links the pattern of trade to factor intensities and factor endowments. A country exports that commodity which uses intensively the factor which is relatively cheap prior to trade.

ii) The *Factor-Price Equalization theorem* suggests that free trade in commodities is sufficient to cause factor prices to be equalized between countries even if factor supplies cannot cross national boundaries.

iii) The *Stolper-Samuelson theorem* states that an increase in the price of some commodity (through tariff policy, e.g.) must unambiguously raise the real reward to some factor of production.

iv) The *Rybczynski theorem* points out that, if prices are kept constant but the endowment of some factor rises, not all outputs can expand. The production of some commodity must fall.

It is convenient to group (i) and (ii) together and (iii) and (iv) together. At a superficial level, the Heckscher-Ohlin theorem and Factor-Price Equalization theorem seem to rely heavily upon an assumption that countries share identical technologies, whereas the Stolper-Samuelson and Rybczynski theorems stress the internal structure of a single economy, quite apart from levels of technology abroad. Indeed this seems especially true of the Factor-Price Equalization theorem, which is a razor's-edge kind of result. If two countries have dissimilar technologies, any basis for an exact equivalence in factor-pricing structure (in the absence of international factor mobility) quite obviously disappears. In similar though less exact fashion, any strong difference in the quality of technologies between countries could outweigh factor endowment differences and upset the Heckscher-Ohlin theorem.

At a deeper level, however, the essence of propositions (i) and (ii) is the link between factor and commodity prices in any particular country. The Heckscher-Ohlin theorem rests upon an assumption that unit costs can be disaggregated into components (labor costs, capital costs) and that changes in factor prices influence unit costs of various commodities differently, depending upon factor proportions. The Factor-Price Equalization theorem, in turn, depends upon the delicate mathematical property of uniqueness of a factor-price vector for a given set of commodity prices for a single economy.[2] The international trade application then follows from the assumption that trade equalizes commodity prices between countries and, of course, that countries share a common technology.

---

[1] These are the same four propositions singled out by Ethier (1974) in a classic paper.

[2] For this reason we do not pursue an analysis of the Factor-Price Equalization theorem in the remainder of the paper. For discussion of this topic, see, e.g., Chipman (1969).

It is important also to point out that the standard Heckscher-Ohlin theory in the simple two-commodity, two-factor case does *not* argue that these theorems *must* hold. Even the simple two-dimensional theory allows these theorems to be violated. For example, if the ranking of commodities by factor intensities does not correspond between countries (the "factor-intensity reversal" phenomenon), the Factor-Price Equalization theorem cannot be satisfied, and the Heckscher-Ohlin theorem must be violated for at least one country.[3] Difficulties of this kind obviously do not go away in higher-dimensional cases. The point that is relevant here, however, is that the assumption of small scale (2 × 2) does not preclude a richer variety of results.

The Stolper-Samuelson and Rybczynski theorems seem more robust. They depend only upon the assumption that each commodity is produced by combining the two factors in proportions that differ between sectors. Each is a different reflection of the *magnification effect* whereby factor prices and commodity output changes are more widespread than commodity price and factor endowment changes, respectively (see Jones 1965*a*, 1965*b*; or Caves and Jones 1973, chaps. 8, 9). Properly qualified, this magnification effect does not depend upon the number of sectors in the economy.

II. Factor Intensities and the Heckscher-Ohlin Theorem

In the 2 × 2 model with labor and capital the two productive factors, one commodity is said to be relatively labor intensive if the ratio of labor to capital employed in producing that commodity is greater than the labor/capital ratio used in the other sector. A double bilateral comparison is involved: a ratio of two factors compared between two industries. Such a concept of factor intensity can be carried over readily to a model incorporating many factors and commodities. Just pick a pair of factors and a pair of commodities, and compare the ratio of factor use for the two industries.

However, it proves more instructive to proceed along different lines. In the 2 × 2 model, alternative but equivalent expressions of the factor-intensity ranking were provided, on the one hand, by a comparison of the distributive share of labor in the two industries (θ_{L1} and θ_{L2}) or, on the other, by a comparison of the fraction of the total supply of each of the two factors employed in the first industry (λ_{L1} with λ_{K1}). If the first industry employs a higher labor/capital ratio than does the second, then θ_{L1}

---

[3] That is, with intensities reversed, either each country exports the commodity which, to it, is labor intensive, or each country exports its capital-intensive good. See Jones (1956).

exceeds θ_{L2}, and, as well, the fraction of the labor force used in the first industry (λ_{L1}) exceeds the fraction of the capital stock employed by the first industry (λ_{K1}). It is a variation on these comparisons rather than physical bilateral proportions that usefully generalizes.

The intensity with which any sector uses a particular factor can be compared with the average for the economy. Let α^i represent factor i's distributive share in the economy as a whole. In contrast, θ_{ij} is factor i's distributive share in the jth sector. Let

$$z_{ij} \equiv \frac{\theta_{ij}}{\alpha^i}, \tag{1}$$

so that, if z_{ij} exceeds unity, the jth sector is said to be "intensive" in its use of factor i. If each z_{ij} were weighted by the importance of factor i in the national income (α^i), the weighted average for *all* factors used in the jth industry,

$$\sum_i \alpha^i z_{ij},$$

would of course be unity.

Along parallel lines, compare the fraction of factor i used in industry j, λ_{ij}, with the fraction of the national income represented by the output of industry j, α_j. (Literally α_j is p_jx_j/Y, where p_j is price, x_j is output, and Y is the value of national income.) Industry j might be termed "intensive" in its use of factor i if λ_{ij} exceeds α_j. Indeed, as it is easy to verify, the ratio λ_{ij}/α_j is equal to z_{ij}.

The Z-matrix thus summarizes information about factor intensities. If an entry, z_{ij}, exceeds unity, industry j uses factor i intensively.[4] The z_{ij} can be compared with any other element in row i or column j. Each column refers to a particular industry's use of each and every productive factor. Thus z_{ij} exceeds z_{hj} if λ_{ij} exceeds λ_{hj}. If so, industry j is intensive in its use of factor i relative to its use of factor h. Such a comparison would be useful in answering the following question: If factor prices are given and the jth output rises (all other outputs constant), does the demand for factor i rise relatively more or less than the demand for factor h? The answer is more, if z_{ij} exceeds z_{hj}.[5]

It is a comparison of elements in a row of the Z-matrix that supports the extension of the Heckscher-Ohlin theorem to a world of many factors

---

[4] Dual interpretations of this concept of intensity are discussed in Section V.

[5] Let a_{ij} denote the amount of factor i used per unit output of commodity j. If factor prices are given, a_{ij} is determined. The total amount of factor i used in the economy, V_i, is

$$\sum_l a_{il}x_l.$$

Therefore, if only output j rises (and factor prices are fixed), $\hat{V}_i = \lambda_{ij}\hat{x}_j$, where ^ over a variable denotes a relative change ($\hat{x}$ is dx/x).

and commodities. The underlying rationale for the Heckscher-Ohlin theorem is the connection between factor prices and unit costs. Unit costs (c_j) are equal to the sum of factor components. Thus

$$a_{1j}w_1 + a_{2j}w_2 + \cdots + a_{rj}w_r = c_j. \tag{2}$$

Taking small changes (and letting $\hat{x}$ be defined as dx/x),

$$\theta_{1j}\hat{w}_1 + \cdots + \theta_{rj}\hat{w}_r = \hat{c}_j. \tag{3}$$

This result presupposes that firms attempt to minimize costs. For any set of factor prices, c_j is minimized when

$$\sum_i \theta_{ij}\hat{a}_{ij}$$

is reduced to zero. Now suppose all we know about two economies before trade is that in the first of them the return to factor 17, w_{17}, is lower than in the other. Then industries can be ranked from highest $\theta_{17,j}$ to lowest $\theta_{17,j}$, and such a factor-intensity ranking would serve as best guide to the extent to which c_j differs in the two countries before trade and thus to the pattern of trade. This of course is only a partial analysis, as differences in other factor prices also affect costs. But the rationale for each factor-price difference between countries affecting unit costs is the same. The impact on costs and the trade pattern depends upon factor-intensity rankings precisely as in the 2 × 2 model.

III. The Magnification Effect: Stolper-Samuelson and Rybczynski Theorems

In the 2 × 2 model, an increase in the price of one commodity causes one factor return to rise relatively more than this price and another factor return to fall. This Stolper-Samuelson version of the magnification effect of commodity-price changes on factor prices holds *regardless* of the number of factors and commodities in the model, as long as there is no joint production. That is, this much-cited feature of the 2 × 2 model is based on a commonly assumed feature of technology: that each productive process combines more than one input to produce only one output. The productive structure is pyramidical (see fig. 1).

No complicated mathematics is required to establish this assertion. Consider equation (3) for all those commodities actually produced, so that unit costs, c_j, represent prices, p_j. Equation (3) states that each price change, $\hat{p}_j$, is a positive weighted average of all factor-price changes. That is, each $\hat{p}_j$ must be trapped between the largest and the smallest of the $\hat{w}_i$'s. Such a relationship holds for every commodity produced in

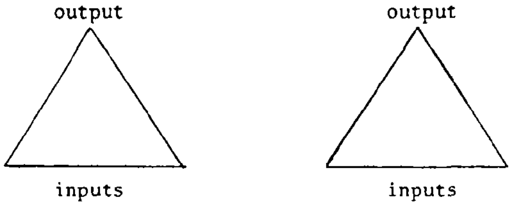


FIG. 1

positive quantities. Thus, if commodities are numbered so that $\hat{p}_1 > \hat{p}_2 > \cdots > \hat{p}_n$, there must exist at least two factors, i and h, such that

$$\hat{w}_i > \hat{p}_1 > \cdots > \hat{p}_n > \hat{w}_h. \tag{4}$$

This proposition is in no way limited to 2 × 2 models.[6]

As a special case, suppose that only one commodity price rises and all other commodity prices remain constant. Then at least one factor of production finds its real reward unambiguously raised, and at least one other factor reward falls. This basic generalization of the Stolper-Samuelson result follows directly from the assumption of no joint production and is independent of the dimensionality assumption as to the number of factors and commodities.[7] Of course we would like to know more *which* factor prices rise and which do not. Answers to these detailed compositional questions depend upon a description of the technology for each and every commodity-factor share and (in cases in which the number of factors exceeds the number of commodities) elasticities of substitution between factors. Chipman (1969), Kemp and Wegge (1969), Inada (1971), Uekawa (1971), and others have investigated special sets of restrictions that allow detailed results in the $n \times n$ case to correspond to those in the 2 × 2 case. The basic "magnification effect" for prices is independent of such restrictions.

If joint production takes the special form in which a single input produces many outputs (see fig. 2), the Stolper-Samuelson type of result would be reversed: the distribution of output prices would be wider than that of input prices. Each input-price change would be a positive weighted

---

[6] In some special cases the strict inequality at the beginning or end of this chain would not hold. The Appendix explains more carefully the conditions required to ensure strict inequalities here and subsequently in the text. As the emphasis on the lack of joint production may suggest, we assume that at least two factors are employed in each productive activity (and, further, that every factor is employed in at least two sectors).

[7] A simple proof for the "even" case ($n \times n$) involving the share matrix and its inverse is provided in Ethier (1974). In these remarks we continue to assume that any commodity whose price changes continues to be produced.

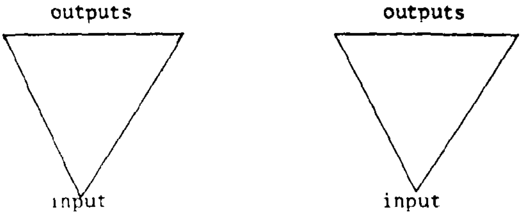


FIG. 2

average of commodity-price changes.[8] Again, it is the asymmetry between inputs and outputs in the productive structure that accounts for the basic theorem establishing asymmetry between the changes in input and output prices.[9]

The magnification effect linking changes in commodity and factor prices is also revealed in the dual ranking of endowment and output changes in the basic 2 × 2 model. The special case in which the endowment of only one factor changes leads to the Rybczynski result: *if prices are constant, the output of some commodity must fall.*[10] But there is a potential ambiguity in the preceding statement of the theorem. Which set of prices is assumed constant? In the 2 × 2 world, commodity and factor prices are *locally* uniquely linked so that, if one set of prices is held constant, so is the other. Fairly weak conditions suffice to establish in the $n \times n$ case that *locally* factor prices and commodity prices are uniquely linked (as opposed to the stronger conditions needed for global factor-price equalization, as in Gale and Nikaido [1965] and Chipman [1969]).

The ambiguity arises in the case in which the number of factors exceeds the number of produced commodities. Suppose commodity prices are fixed in such a model. Let the endowment of some factor rise. This results in an outward shift in the transformation surface and (at the given commodity prices) a unique change in outputs. But factor prices are also disturbed and with them techniques. Thus holding commodity prices constant is different from holding factor prices (and all techniques) constant.

---

[8] There is a branch of economics that employs the basic assumption that each input produces many outputs. This is the "new" theory of demand associated with the work of Lancaster (1971), in which each commodity (input) provides a variety of characteristics (outputs) of value to consumers.

[9] Joint production may not involve such asymmetries: many inputs may be required to produce, jointly, many outputs. But an asymmetry can be created by focusing on the concept of a production process (one) involving many inputs and many outputs. Indeed, a central issue for trade theory would be the question of where the *processes* are located and how factor intensities and factor endowments are instrumental in determining location patterns.

[10] The basic reference is Rybczynski (1955). See also Jones (1956). Duality to the Stolper-Samuelson result is discussed in Jones (1965*a*).

To see why this is an important issue, consider the basic relationship between outputs and a typical factor endowment (V_i), if that factor is fully employed:

$$a_{i1}x_1 + a_{i2}x_2 + \cdots + a_{in}x_n = V_i. \tag{5}$$

Let the initial equilibrium be disturbed by a change in factor endowments, with no restriction on factor prices. If full employment can still be maintained,

$$\lambda_{i1}\hat{x}_1 + \lambda_{i2}\hat{x}_2 + \cdots + \lambda_{in}\hat{x}_n = \hat{V}_i - \left(\sum_j \lambda_{ij}\hat{a}_{ij}\right). \tag{6}$$

Each λ_{ij} represents the fraction of factor i that is employed in the jth industry. If factor prices are kept constant, the term

$$\sum_j \lambda_{ij}\hat{a}_{ij}$$

disappears as techniques are unaltered. If such is the case, the form of (6) looks much like that of (3). Each endowment change, $\hat{V}_i$, would be a positive weighted average of all the output changes, and this would once again reflect the assumption of no joint production.

These remarks suggest that, if the factor price commodity price link is not unique, the generalization of the Rybczynski theorem should be phrased in terms of constant factor prices. This would, we feel, be inappropriate for the following two reasons. First, if the number of factors exceeds the number of commodities and *if factor prices are held constant*, relationships such as (6) cannot hold for all factors. Even if all factors are initially fully employed, they cannot remain so unless all factor endowments expand in proportion. Second, the remarkable *reciprocity* relationship shown in (7), which serves as the basis

$$\frac{\partial w_i}{\partial p_j} = \frac{\partial x_j}{\partial V_i} \tag{7}$$

for the dual relationship between the Stolper-Samuelson and Rybczynski results, assumes (in the $\partial x_j/\partial V_i$ term) that all *commodity prices* are held constant.[11] We postpone until Section V a proof of this relationship.

The case in which the number of commodities exceeds the number of factors is different. If commodity prices are arbitrarily chosen, with no link to local technology, the economy would tend to specialize to produce no more commodities than factors. On the other hand, a closed economy consuming and producing all commodities has its commodity prices reflecting local costs. For example, consider a Ricardian (one-factor) economy producing n goods on the interior of a flat transformation surface. Suppose that the labor endowment expands but commodity prices stay constant. Without information about tastes, of course, output changes are not uniquely determined. But for *any* feasible output changes it must be the case that the relative change in the labor endowment is a positive weighted average of the relative changes in outputs. Outputs either all rise in the same proportion as labor, or some rise more and others less. If there should be more than one factor (but still many more goods) and the endowment of one of them does not change, at least one output falls.

---

[11] This relationship was stated and proved by Samuelson (1953).

IV. "Even" versus "Uneven" Cases

The 2×2 model is "even"—the same number of factors and commodities. Generalizations to higher dimensions need to account not only for "even" $n \times n$ versions but also for cases in which the number of commodities, n, differs from the number of productive factors, r. Given the tendency to specialize in open economies if $n > r$, here we restrict ourselves to the "uneven" case in which $r > n$.

The Stolper-Samuelson generalization discussed in the preceding section was not restricted to the $n \times n$ case. A relationship like (4) reflects the absence of joint production. For example, if just one commodity price, p_j, rises, some factor price must rise in greater proportion, and some other factor price must fall. But must each factor in turn play these extreme roles for some commodity-price changes?

In examining this question, consider a simple version of an "uneven" case. It is the "specific-factors" model or what Samuelson has termed the "Ricardo-Viner" model.[12] Restrict the number of commodities to two, and suppose that one sector uses capital and labor and the other uses land and labor. Capital and land are each specific to one sector, while labor is mobile—it can be used in either sector and earns the same return in each. Now consider how the wage rate is affected by a rise in either commodity price. Labor is attracted to whichever sector benefits from the price rise. Therefore labor's marginal product must rise in the sector in which price is constant (since the ratio of labor to the specific factor there falls). The wage rate must therefore rise. But in real terms the wage rate falls in the sector where price has risen. To be explicit, if $\hat{p}_1 > \hat{p}_2$,

$$\hat{p}_1 > \hat{w} > \hat{p}_2.$$

Labor plays an "intermediate" role. It is never the extreme factor.

In the even $n \times n$ case, with perhaps all factors mobile from sector to sector, this kind of behavior cannot arise. In particular, the following result can easily be proved: *if the number of factors equals the number of*

---

[12] This model is discussed in Jones (1971, 1975), Samuelson (1971), Mayer (1974), and Mussa (1974).

commodities, every factor has at least one natural enemy. That is, pick a factor i. It is always possible to find a commodity, j, such that, if the price of j rises and no other commodity prices change, w_i must fall. In this sense industry j is a "natural enemy" of factor i.

The proof is simple and makes use of the reciprocity relationship, (7). In the "even" case, ask what happens if, at constant commodity prices, the endowment of factor i, V_i, should expand, all other endowments constant. Because $n = r$, factor prices and techniques are also constant. From relationships like (6), simplified now with every $\hat{a}_{ij}$ zero, we know that for some j, k,

$$\hat{x}_k > \hat{V}_i > 0 > \hat{x}_j. \tag{8}$$

Each factor endowment change is trapped between output changes. Since all endowments other than V_i are constant, some x_j must fall. The reciprocity relationship, (7), then ensures that, if commodity j experiences a price rise (all other prices constant), the return to factor i would fall. In the even case, every factor has at least one natural enemy. And, as the specific factors model reveals, this is a peculiarity of even cases not necessarily shared by the uneven cases.

If in the even case every factor has a natural enemy, does each factor as well have at least one unambiguous natural friend? That is, for each factor, i, is it possible to find at least one industry, k, for which $\hat{w}_i > \hat{p}_k$ if p_k rises with all other commodity prices constant? Can dual relationship (8) be used to establish this? No. What (8) proves is that for each factor, i, there is some industry, k, such that, if V_i rises (all other endowments constant),

$$\frac{\partial x_k}{\partial V_i} > \frac{x_k}{V_i}.$$

Reciprocity relationship (7) identifies $\partial x_k/\partial V_i$ with $\partial w_i/\partial p_k$. Thus

$$\frac{\partial w_i}{\partial p_k} > \frac{x_k}{V_i},$$

or, in relative terms,

$$\frac{\hat{w}_i}{\hat{p}_k} > \frac{\alpha_k}{\alpha^i},$$

where, as before, α^i is factor i's share of the national income and α_k is the share of commodity k's output in the national income.[13] If factor i is *unimportant* enough, in the sense of a low factor share, α^i, compared with α_k, then indeed commodity k is a natural friend in the sense that $\hat{w}_i$

---

[13] For a discussion of the relationship between Stolper-Samuelson terms such as $\hat{w}_i/\hat{p}_k$ and Rybczynski terms such as $\hat{x}_k/\hat{V}_i$, see Ethier (1974) and Ruffin and Jones (1977).

would exceed $\hat{p}_k$. But no general proof along these lines can be constructed.

To establish this, consider the following counterexample.[14] Let

$$\theta' = \begin{bmatrix} \frac{8}{10} & 0 & \frac{2}{10} \\ \frac{7}{20} & \frac{7}{20} & \frac{6}{20} \\ 0 & \frac{6}{15} & \frac{9}{15} \end{bmatrix}$$

be the matrix of distributive shares in a case of three factors and three commodities. Elements in each row refer to the factor shares in a particular industry and thus must add to unity. The inverse of this matrix is

$$(\theta')^{-1} = \begin{bmatrix} \frac{9}{10} & \frac{4}{5} & -\frac{7}{10} \\ -\frac{21}{10} & \frac{24}{5} & -\frac{17}{10} \\ \frac{7}{5} & -\frac{16}{5} & \frac{14}{5} \end{bmatrix}.$$

Each row shows what would happen to a particular factor price if each commodity price, in turn, were raised. For example, the $\frac{7}{5}$ entry in the third row, first column, shows that, if the price of the first commodity were to rise by 10 percent, other commodity prices constant, the return to the third factor would rise by 14 percent. Now focus on the first row. No entry exceeds unity. It is not possible to find an industry for which protection (a price rise) would unambiguously raise the real return to the first factor. *Some* return would have to rise—for example, that of the second factor if the second price rises. Note also that each row has a negative entry—every factor has a natural enemy.

In the original literature, the Stolper-Samuelson result aimed at establishing an unambiguous winner when a commodity price rises. There will always be at least one, but, as shown above, not every factor can rely upon some commodity-price rise (or combination of price rises) to improve its real position. Thus carte blanche in imposing tariffs may not prove sufficient to raise the real wage of some factors, even in the $n \times n$ (even) case.

In contrast, consider the Rybczynski result. It purported to show some output *falling* when the endowment of some factor increases, other endowments (and commodity prices) held constant. This must be the case if the number of factors equals the number of commodities—this was shown by (8). But if factors exceed commodities in number, no output need fall as one endowment rises. The specific-factors model again provides the illustration: at constant prices an increase in mobile labor causes all outputs to rise. Nevertheless, the following Rybczynski result does generalize: for each output, j, it is possible to find a factor of production, i, such that if V_i increases (other V's and p's constant) x_j must fall. In this sense each *industry* has a natural enemy. *Proof:* for each commodity, j, there is some factor, i, such that w_i falls if p_j rises. Using the reciprocity relationship(7), an increase in V_i must reduce x_j. The number of factors may exceed the number of commodities.

---

[14] This matrix appears in Uekawa (1971).

V. General Properties of Production Models

Some properties of the 2 × 2 model hold in a wide variety of cases, including both even and uneven cases (still assuming $r \geq n$) with large numbers of factors and commodities as well as cases involving joint production. To conclude this paper we consider three such properties, the last of them the basic reciprocity relationship already stated in equation (7).

The appropriate degree of generality is attained by considering a set of productive activities. The unit level of activity j involves a set of inputs, a_{ij}, and a set of outputs, b_{kj}. The competitive profit conditions maintain that in equilibrium unit costs cannot fall short of the value of operating the activity at unit level. Thus

$$\sum_i a_{ij} w_i \geq \sum_k b_{kj} p_k. \tag{9}$$

If actual losses are made, the level of the activity, s_j, would be reduced to zero. For activities involving positive production, unit costs are reduced to a minimum by appropriate choice of techniques. This entails

$$\sum_i w_i \, da_{ij} - \sum_k p_k \, db_{kj} = 0. \tag{10}$$

Therefore, differentiating (9), we obtain

$$\sum_i a_{ij} \, dw_i \gtreqless \sum_k b_{kj} \, dp_k. \tag{11}$$

The equality sign holds if production level s_j is positive. The inequalities do not disturb matters, since each relationship in (11) is now multiplied by the activity level. Summing over all j,

$$\sum_j s_j \sum_i a_{ij} \, dw_i = \sum_j s_j \sum_k b_{kj} \, dp_k, \tag{12}$$

or

$$\sum_i V_i \, dw_i = \sum_k x_k \, dp_k. \tag{13}$$

This can also be written in relative terms as

$$\sum_i \alpha^i \hat{w}_i = \sum_k \alpha_k \hat{p}_k, \tag{14}$$

where the α^i are input shares in the national income and the α_k are output shares.

Expression (14) states a widespread property of any general equilibrium model in which monetary factors are excluded so that the prive *level* has no real impact. It accommodates an equal proportional change in all output and input prices.

The second general property is similar to (14) except in involving physical outputs and inputs instead of prices. It states,

$$\sum_i \alpha' \hat{V}_i = \sum_j \alpha_j \hat{x}_j. \tag{15}$$

Assuming there is no technical progress, (15) asserts that any changes in factor endowments are matched, in the aggregate, by a change in physical outputs. The weights are once again the input and output shares in the national income. Relationship (15) is derived by considering the equality between the two ways of expressing national income:

$$\sum w_i V_i = \sum p_j x_j. \tag{16}$$

Differentiate (16) and cancel the price variations by using (14) to obtain (15).

A special case of (15) involves a general statement of the marginal productivity relationship. Suppose only one factor endowment, V_i, changes. Then (15) can be written as

$$w_i = \left(\sum_i p_j \, dx_j\right) \Big/ dV_i. \tag{17}$$

This identifies the competitive return to factor i as the change in the value of aggregate output at initial prices that would accompany an increase in V_i of one unit. The remarkable feature of (17) is that prices are not necessarily being held constant, nor are techniques of production necessarily smoothly flexible, nor indeed is joint production ruled out.

Section II of this paper alluded to dual interpretations of the concept of factor intensity embodied in the Z-matrix where z_{ij} referred either to θ_{ij}/α' or to λ_{ij}/α_j.[15] Industry j was said to use factor i intensively if z_{ij} exceeded unity. This involves a comparison of j with the economy as a whole. Suppose that endowments are held constant but w_i is raised by 1 percent. As (14) suggests, the aggregate price *level* is in this case raised by the fraction α' of 1 percent. In contrast, single price p_j is raised by the fraction θ_{ij} of 1 percent. Industry j is deemed to be intensive in its use of factor i if a given change in w_i has a relatively stronger impact on the price of commodity j than it does on the aggregate price level.

The dual interpretation of z_{ij} utilizes (15). Suppose the economy wishes to expand output in the jth sector by 1 percent, without reducing output in any other sectors and without changing techniques. This would require an increase in the economy's factor endowments. As (15) states, the output share, α_j, would express the fraction of 1 percent by which demand for all productive factors rises. The λ_{ij}, the fraction of V_i used in the jth industry, expresses the fraction of 1 percent by which demand for factor i is increased when x_j expands by 1 percent. If z_{ij} is greater than unity, an expansion of j's output increases demand for factor i by a greater relative amount than it does for resources on average. In this sense j is factor-i intensive.

---

[15] We assume now that there is no joint production.

Both properties (14) and (15) are crucial in establishing the basic reciprocity relationship captured by (7). National income, Y, is expressed either as the sum of factor payments,

$$\sum_h w_h V_h,$$

or as the aggregate value of output,

$$\sum_k p_k x_k.$$

Differentiate the latter expression with respect to p_j, holding all other commodity prices and all factor endowments constant. This yields

$$\frac{\partial Y}{\partial p_j} = x_j + \sum p_k \frac{dx_k}{dp_j}.$$

Such a price change in general causes outputs to respond along the transformation schedule. But relationship (15) states that, if all endowments are fixed, the aggregate value of output changes at initial prices, $\sum p_k\, dx_k$, must vanish. Therefore

$$\frac{\partial Y}{\partial p_j} = x_j. \tag{18}$$

In similar fashion, differentiate Y expressed as

$$\sum_h w_h V_h$$

with respect to V_i, holding all commodity prices and all other factor endowments constant. Thus

$$\frac{\partial Y}{\partial V_i} = w_i + \sum_h V_h \frac{dw_h}{dV_i}.$$

In special cases (the $n \times n$ case) each factor-price change is zero because commodity prices are being held constant. But in any event the sum

$$\sum_h V_h\, dw_h$$

must vanish by (14) if each commodity price is held fixed. Thus

$$\frac{\partial Y}{\partial V_i} = w_i. \tag{19}$$

The reciprocity relationship, (7), follows by differentiating (18) with respect to V_i and (19) with respect to p_j and equating the second partials of income with respect to p_j and V_i regardless of the order of differentiation. This relationship can also be used to link *relative* changes in factor and commodity prices (the Stolper-Samuelson terms) on the one hand and *relative* changes in outputs and endowments (the Rybczynski terms) on the other:[16]

$$\alpha^i \cdot \frac{\hat{w}_i}{\hat{p}_j} = \alpha_j \frac{\hat{x}_j}{\hat{V}_i}. \tag{20}$$

This relationship underlies the asymmetry established in the preceding section: in the $n \times n$ case each factor had a "natural enemy," since a negative $\hat{x}_j/\hat{V}_i$ implies a negative $\hat{w}_i/\hat{p}_j$. But each factor need not, in general, have a "natural friend," since a value of $\hat{x}_j/\hat{V}_i$ greater than unity is no guarantee that $\hat{w}_i/\hat{p}_j$ also exceeds unity. Factor and output shares are also involved.

VI. Concluding Remarks

The two-sector model of production as used in the theory of international trade has a rich variety of properties. This paper has shown that relatively few of them are directly dependent upon the small dimensionality represented by two commodities and two factors. The concept of factor intensity and the rationale of the Heckscher-Ohlin theorem generalize readily to higher-dimensional cases. The crucial assumption supporting the Stolper-Samuelson and Rybczynski theorems is the absence of joint production. Identifying the detailed composition of output and factor-price changes is difficult, in higher dimensions, but the existence of magnification effects, especially in the even case in which the number of factors equals the number of goods, is easy to establish. Throughout, the argument has deliberately been kept free of formal mathematics. The Appendix provides an alternative formal treatment of the general model.

Mathematical Appendix

Let technology be summarized by an $r \times n$ nonnegative matrix of input-output coefficients, A. The number of factors, r, is greater than or equal to the number of commodities, n. Each element of A, a_{ij}, is assumed to be a unique zero homogeneous function of all factor prices:

---

[16] This relationship is also discussed in Ruffin and Jones (1977). Note that endowments and all p_k, $k \neq j$, are held constant in $(\hat{w}_i/\hat{p}_j)$ and all commodity prices and all V_h, $h \neq i$, are held constant in $(\hat{x}_j/\hat{V}_i)$.

$$a_{ij} = a_{ij}(w_1, \ldots, w_r). \tag{A1}$$

We assume this can be differentiated to yield

$$da_{ij} = \sum_{k=1}^{r} a_{ij}^{k}\, dw_k, \tag{A2}$$

where a_{ij}^{k} is defined as $\partial a_{ij}/\partial w_k$. Of course each a_{ij}^{i} is negative (or zero), and

$$\sum_k w_k a_{ij}^{k} = 0.$$

Suppose that the technology possesses sufficient flexibility to allow full employment. With joint production ruled out and with x representing the vector of outputs and V the vector of factor endowments, the full-employment conditions can be stated as

$$Ax = V. \tag{A3}$$

Consider the ith factor. For small changes,

$$\sum_j x_j\, da_{ij} + \sum_j a_{ij}\, dx_j = dV_i. \tag{A4}$$

Substitute (A2) for da_{ij} to obtain

$$\sum_k s_{ik}\, dw_k + \sum_j a_{ij}\, dx_j = dV_i, \tag{A5}$$

where s_{ik} is defined as

$$s_{ik} \equiv \sum_j x_j a_{ij}^{k}.$$

The s_{ik} shows how the economy's demand for factor i would respond at unchanged outputs as the return to factor k rises. Each s_{ii} is negative (or zero), and

$$\sum_k w_k s_{ik} = 0.$$

In matrix notation, the differentiated full-employment conditions can be written as

$$S\, dw + A\, dx = dV, \tag{A6}$$

where S is the $r \times r$ matrix $[s_{ik}]$, dw the r-element column vector (dw_i), and dx the n-element column vector (dx_j).

The dual competitive profit conditions, assuming that all industries operate at positive levels, are shown by

$$A' \cdot w = p. \tag{A7}$$

With technology sufficiently flexible to yield full employment, we now assume that all factor prices and commodity prices are strictly positive. Cost minimization in each industry ensures that

$$\sum_i w_i\, da_{ij} = 0,$$

so that the equations of change, in matrix notation, are

$$A' \cdot dw = dp. \tag{A8}$$

It proves convenient to write equation sets (A6) and (A8) together as

$$\left[\begin{array}{c|c} S & A \\ \hline A' & 0 \end{array}\right] \cdot \left[\begin{array}{c} dw \\ \hline dx \end{array}\right] = \left[\begin{array}{c} dV \\ \hline dp \end{array}\right]. \tag{A9}$$

Properties of this partitioned matrix and its inverse are discussed below. For easy reference, let

$$B \equiv \left[\begin{array}{c|c} S & A \\ \hline A' & 0 \end{array}\right]. \tag{A10}$$

The substitution matrix S is symmetric and negative semidefinite. We will assume that we have a regular minimum of the cost function (see Samuelson 1947, p. 68); that is, the corresponding substitution matrix S^j for each commodity satisfies $u'S^ju < 0$ for every nonzero u not proportional to w.[17] Hence we have

$$u'Su < 0 \qquad \text{for } u \text{ not proportional to } w,\ u \neq 0. \tag{A11}$$

In particular, S has rank $r - 1$ and a negative diagonal, since w is a strictly positive vector.

Also, since p is a positive vector, (A7) implies that, if $u \neq 0$ and $A'u = 0$, then u is not proportional to w. Hence, if $u \neq 0$ and $A'u = 0$, by (A11), $u'Su < 0$. Hence B is a bordered matrix that represents a quadratic form which is negative definite under constraints. In particular, if $r \geq n$, B has full rank $n + r$, since the rank of A is n (see Caratheodory 1967, pp. 195–96; and Samuelson 1947, pp. 378–79). Its inverse is written as in (A12):

$$B^{-1} \equiv \left[\begin{array}{c|c} S & A \\ \hline A' & 0 \end{array}\right]^{-1} \equiv \left[\begin{array}{c|c} K & G \\ \hline G' & L \end{array}\right] \tag{A12}$$

The inverse of B has some known properties (see, once again, Caratheodory 1967; and Samuelson 1947). In particular, K is symmetric and negative semidefinite with rank $r - n$.

Solving for (dw, dx), we obtain

$$\left[\begin{array}{c} dw \\ \hline dx \end{array}\right] = \left[\begin{array}{c|c} K & G \\ \hline G' & L \end{array}\right] \left[\begin{array}{c} dV \\ \hline dp \end{array}\right]. \tag{A13}$$

It is clear that the K-matrix answers the following question: What is the impact of a change of factor endowments on factor prices if commodity prices are kept constant? Since K has rank $r - n$, it is clear that in the even case, $n = r$, the K-matrix reduces to a zero matrix. This is the factor-price equalization result. But if $r > n$, factor endowments would have an impact on factor prices. However, it is not true that K must have a negative diagonal. That is, even with $r > n$, an increase in a factor endowment does not have to lower (although it cannot raise) that factor's price at constant commodity prices. For instance, if a good uses only one factor of production, that factor's price cannot change no matter what happens to its endowment, since the commodity price is being held constant.

---

[17] This restriction on the values of u is necessary because the homogeneity property requires that

$$\sum_k w_k s_{ik} = 0.$$

More about K can be learned by writing explicitly the equation $BB^{-1} = I$, as in (A14):

$$\left[\begin{array}{c|c} S & A \\ \hline A' & 0 \end{array}\right] \cdot \left[\begin{array}{c|c} K & G \\ \hline G' & L \end{array}\right] = \left[\begin{array}{c|c} I & 0 \\ \hline 0 & I \end{array}\right]. \tag{A14}$$

Notice that $0 = A'K \equiv KA$. Hence all vectors of endowment changes which are linear combinations of columns of A leave factor prices unchanged.[18] Furthermore, since the rank of K is $r - n$ and the rank of A is n, those are the only vectors of endowment changes that leave factor prices unchanged. Hence (A15) holds:

$$Ku = 0 \qquad \text{iff } u = Aq \text{ for some } n\text{-vector } q. \tag{A15}$$

It follows that, *if every* r − 1 *rows of* A *contain* n *linearly independent vectors, then the diagonal elements of* K *are negative*. To see this, observe that, since K is symmetric and negative semidefinite, if $u'Ku = 0$, then $Ku = 0$. Let e_i denote the r-vector whose ith coordinate is 1, and all other coordinates are zero. If $K_{ii} = e_i'Ke_i = 0$, then $Ke_i = 0$, and by (A15) there exists $q \neq 0$ such that $Aq = e_i$. Such q is then perpendicular to every row of A except the ith row. By assumption, there exist n linearly independent n-vectors among those $r - 1$ rows of A. Hence q is an n-vector perpendicular to n linearly independent n-vectors, and hence $q = 0$. But this contradicts $Aq = e_i$.

If the condition stated above is not satisfied, then a change in the endowment of one factor may not affect that factor's price. To see this, consider

$$A = \begin{bmatrix} 1 & 2 \\ 1 & 1 \\ 1 & 1 \end{bmatrix}.$$

If $q = (-1, 1)$, then $Aq = e_1$. Hence $Ke_1 = KAq = 0$, since $KA = 0$. Hence a change in the first factor endowment does not change any factor prices. But notice that in this case, if we leave out the first row of A, we obtain a matrix of rank 1, and hence the hypothesis of the theorem is not satisfied. This illustrates the importance of looking at the rows of A, that is, how each particular factor enters the input-output matrix.

A simple diagram can point out the importance of the rows of A. Suppose that all factors are fully employed in a three-factor, two-commodity model. The ith row of A shows how the two outputs are constrained by the limitation of factor V_i. This constraint is linear, and for full employment all three constraints must intersect at the same point, as in figure 3. Clearly a change in any single factor endowment would no longer entail a common intersection point. Therefore factor prices would have to change. However, suppose that rows 2 and 3 of A were linearly dependent. In such a case the V_2 and V_3 constraint lines in the diagram would be parallel—and full employment would necessitate their lying on top of each other. A change in the endowment of factor 1 could thus be accommodated by output changes without the necessity of any change in factor prices. Note also that, if all endowments expand proportionally to the input requirements in x_1,

---

[18] Suppose that endowment changes are proportional to the input requirements for commodity j. Then at constant commodity prices these endowment changes can be exactly absorbed by changes in j's output without requiring other outputs to adjust or techniques (through changes in factor prices) to be altered. To see that only j's output changes, note that (A14) provides that $A'G = I = G'A$ and that, by (A13), $dx = G'dV$ at constant commodity prices. Therefore, if dV (nonnegative) is proportional to the jth column of A, dx_j is positive, and $dx_k = 0$ for all $k \neq j$.

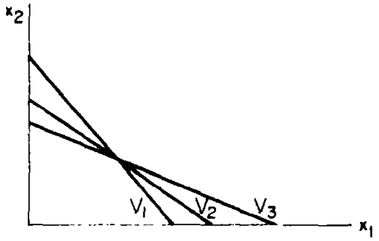


FIG. 3

the common intersection point shifts to the right—an increase in x_1 output alone can maintain full employment without any change in factor prices.[19]

The $n \times n$ submatrix L relates output changes along a given transformation surface to commodity-price changes. From (A14), we have

$$SG + AL = 0.$$

Therefore

$$AL = -SG,$$

or

$$G'AL = -G'SG.$$

But, when the second row of the B-matrix is multiplied by the second column of B^{-1}, we obtain

$$A'G = I, \tag{A16}$$

or, what is the same thing,

$$G'A = I.$$

Therefore

$$L = -G'SG.$$

Pre- and postmultiply by a vector, x, to obtain

$$x'Lx = -(Gx)'S(Gx). \tag{A17}$$

The negative semidefiniteness of S establishes that L is a *positive semidefinite* matrix. Using the fact that the rank of S is assumed to be $r - 1$, one can show that L has a *positive* diagonal whenever $n \geq 2$. For if $L_{ii} = 0$, we have (using [A17]) $0 = -(Ge_i)'S(Ge_i)$. By (A11) this means that $Ge_i = \alpha w$ for some real number α. However, (A16) and (A17) (together with [A7]) imply that $e_i = A'Ge_i = A'\alpha w = \alpha p$. If $n \geq 2$, this means that some coordinate of the price vector is zero, which contradicts our assumptions. Thus, as we would expect, a rise in any commodity price must increase the output of that commodity along the (strictly bowed out) transformation surface.[20]

Note the roles of G and G', as shown by (A13). The G-matrix reveals how a change in commodity prices affects factor prices as of given factor endowments. The G'-matrix links endowment changes (at constant commodity prices) to output changes. That these matrices are indeed the transposes of each other confirms the Samuelson *reciprocity* relation. It proves convenient to alter the B-matrix so that the relationship between $\hat{V}$ and $\hat{p}$, on the one hand, and $\hat{w}$ and $\hat{x}$, on the other, can be displayed. Both the Stolper-Samuelson and the Rybczynski results deal in these *relative* changes.

---

[19] See the discussion in n. 18 above.

[20] If $n = 1$ and factor endowments are constant, output cannot respond to price. A positive diagonal for L presupposes the existence of at least one other sector from which resources can be drawn when the price of some given sector rises.

Return to the basic equation set (A9). Rewriting in terms of relative changes yields

$$\left[\begin{array}{c|c} \sigma & \lambda \\ \hline \theta' & 0 \end{array}\right] \cdot \left[\begin{array}{c} \hat{w} \\ \hline \hat{x} \end{array}\right] = \left[\begin{array}{c} \hat{V} \\ \hline \hat{p} \end{array}\right]. \tag{A18}$$

The σ_{ij} is the element $(w_j \cdot s_{ij})/V_i$. It is a substitution elasticity, showing the percentage of change in the economy's use of factor i (at unchanged outputs) associated with a 1 percent increase in the jth factor price. The σ has a negative diagonal and zero row sums. The λ is the $r \times n$ matrix of factor-allocation coefficients. Thus λ_{ij} is $a_{ij}x_j/V_i$, the fraction of the ith factor employed in the jth industry. The λ is a nonnegative row-stochastic matrix. The θ' is $n \times r$, a matrix of distributive shares. The θ_{ij} is, of course, $w_i a_{ij}/p_j$. The θ' is also row stochastic, so that the entire matrix is row stochastic. We are primarily concerned with the properties of the inverse matrix:[21]

$$\left[\begin{array}{c} \hat{w} \\ \hline \hat{x} \end{array}\right] = \left[\begin{array}{c|c} M & N \\ \hline Q & R \end{array}\right] \left[\begin{array}{c} \hat{V} \\ \hline \hat{p} \end{array}\right]. \tag{A19}$$

The M- and R-matrices are closely related to the K- and L-matrices already discussed. They serve to put factor endowment–factor price and commodity price–commodity output relationships into elasticity form. Stolper-Samuelson results are displayed by N and Rybczynski results by Q. It is to these we now turn. We assume at the outset sufficient richness in the technology so that every factor is used in positive quantities by at least two industries and every industry employs positive amounts of at least two factors.

Consider what can be said in general terms about N and Q by assuming $r \geq n$ and writing explicitly,

$$\left[\begin{array}{c|c} \sigma & \lambda \\ \hline \theta' & 0 \end{array}\right] \cdot \left[\begin{array}{c|c} M & N \\ \hline Q & R \end{array}\right] = \left[\begin{array}{c|c} I & 0 \\ \hline 0 & I \end{array}\right]. \tag{A20}$$

Multiply the second row of the first matrix by the second column of the second to obtain

$$\theta' \cdot N = I. \tag{A21}$$

Ethier (1974) used this relationship in the $n \times n$ case (in which also all a_{ij} were assumed strictly positive) to establish that every column of the N-matrix has at least one negative element. His method of proof extends more generally to the $r \geq n$ case.

Consider the jth column of N. By (A21) this column must be nonzero, so that there exists some k such that $N_{kj} \neq 0$. If the jth column of N is premultiplied by any ith row of θ', (A21) reveals the product to be zero as long as $i \neq j$. We wish to choose row i so that $\theta'_{ik} > 0$. This is possible since every factor k, is by assumption used in at least two industries, and even if one of these should be industry j there will exist some $i \neq j$ for which θ'_{ik} is positive. Therefore $\theta'_{ik}N_{kj} \neq 0$. Now N_{kj} is either negative or positive. If $N_{kj} > 0$, there must exist at least one other element of the jth column of N, N_{hj}, $h \neq k$, such that $N_{hj} < 0$, since

---

[21] The existence of the inverse is obviously guaranteed by the invertibility of B.

$$\sum_k \theta'_{ik} N_{kj} = 0 \qquad \text{for } i \neq j.$$

Therefore at least one element in every column of N must be negative.

Ethier was also able to establish that every column of N must contain at least one element greater than unity. Here his assumption that all a_{ij} are strictly positive turns out to be crucial. *Premultiplying the jth column of N by the jth row of θ'* yields unity. Therefore, if all θ'_{jk} are positive, at least one N_{kj} must exceed unity, since at least one other element in the jth column of N has been proved to be negative, and θ' is row stochastic. The following 3 × 3 example shows wherein this result can break down if all factors are not assumed to be positively employed in all industries. Suppose that

$$\theta' = \begin{bmatrix} 0 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & 0 \end{bmatrix}.$$

In this case, the inverse is

$$N = \begin{bmatrix} -1 & 1 & 1 \\ 1 & -1 & 1 \\ 1 & 1 & -1 \end{bmatrix}.$$

Although every column of N has at least one negative element, no column of N has any elements greater than unity. The jth column of N displays the relative impact on all factor prices of a change in the price of commodity j with all other commodity prices (and factor endowments) constant. If factor j is not used in industry j, an increase in p_j can be passed on uniformly to all factors $k \neq j$, as in this example, without driving any w_k up more than proportionally to p_j. The row-stochastic property of θ' guarantees that either the jth column of N has an element exceeding unity *or* the factor whose wage goes down when p_j rises (and there must be at least one such factor which loses) is not used in the production of the jth commodity.[22]

Must properties of this type characterize the *rows* of N when $r > n$? No. The paper described the specific-factors model as an illustration in which some row of N (the one showing the impact of each and every commodity-price rise on the wage rate) consisted entirely of positive fractions.

When the number of commodities *equals* the number of factors (the even case), more can be said about the rows of N. Rewrite (A20) as

$$\left[\begin{array}{c|c} M & N \\ \hline Q & R \end{array}\right] \left[\begin{array}{c|c} \sigma & \lambda \\ \hline \theta' & 0 \end{array}\right] = \left[\begin{array}{c|c} I & 0 \\ \hline 0 & I \end{array}\right]. \tag{A22}$$

In the even case, the M-matrix is the zero matrix—the local factor-price equalization result. Therefore multiplying the first row of the first matrix by the first column of the second yields

$$N\theta' = I.$$

Consider now the ith *row* of N. Postmultiplying by the kth column of θ' ($k \neq i$) yields the result that every *row* of N must contain at least one negative element, since the elements of θ' are nonnegative.[23] The ith *row* of N describes what happens to w_i as, in turn, each commodity price rises. At least one commodity-price rise must thus lower w_i. Each factor has at least one natural "enemy" in the even ($n = r$) case. Note that multiplying the ith row of N by the ith column of θ' yields an inner product of unity. But this cannot be used to prove that some element of the ith row of N must exceed unity, because the ith *column* of θ' does not add to unity. The ith column of θ' refers to the shares of factor i in all the industries. If factor i is an "important" factor, little can be said, as the ith column would add to greater than unity.[24] But if it is unimportant (with column sum less than unity), then at least one element of the ith row of N must exceed unity. As suggested in the main body of the paper, a relatively unimportant factor could unambiguously be protected by some set of commodity price increases. *But it is not necessarily the case that the "important" factors each have a natural "friend."*

---

[22] Note that, in the 2 × 2 case, the assumption that every factor is used in at least two industries is sufficient to establish a positive θ' matrix.

The Q-matrix, whose elements reveal the Rybczynski results, is not as strongly structured as is the Stolper-Samuelson N-matrix. This difference, to be developed below, is a basic reflection of the asymmetry in the two equation sets captured by (A18). The competitive profit conditions assume the simple form:

$$\theta'\hat{w} = \hat{p},$$

because of the envelope result expressed by the cost-minimization condition $\sum w_i \, da_{ij} = 0$. In contrast, the full-employment conditions are more complicated:

$$\sigma\hat{w} + \lambda\hat{x} = \hat{V}.$$

Even with commodity prices constant, factor prices can change in the case $r > n$.

If the number of factors exceeds the number of produced commodities, it is possible to prove that each *row* of Q has at least one negative element but not that it contains an element greater than unity. Nothing in general can be said about the columns of Q. To verify these remarks, multiply the second row of the first matrix in (A22) by the second column of the second matrix to obtain

$$Q\lambda = I.$$

The inner product of the ith row of Q with the kth column of λ ($k \neq i$) is zero. Since all elements in the λ-matrix are nonnegative, some element in the ith row of Q must be negative.[25]

The elements in the *columns* of λ need not add to unity. The ith column of λ refers to the percentage of each and every factor used in the ith sector. If output i bulks large in the national income, the sum

$$\sum_h \lambda_{hi}$$

---

[23] Here we make use of the assumption that every row of θ' must contain at least two positive elements, i.e., that every good uses at least two distinct factors.

[24] As the preceding example illustrated (for the columns of N), having a column of θ' add exactly to unity here might not suffice to establish the existence of an element in the row of N actually exceeding unity. Formally, we define factor i as *important* if

$$\sum_i \theta'_{ij} = \sum_j \theta_{ij} \geq 1.$$

[25] The proof follows the lines used earlier in establishing the properties of the N matrix. It is necessary to assume, as we have, that each factor is used in the production of at least two commodities.

will greatly exceed unity, and it would not be possible to prove that any element of the ith row of Q would have to be larger than one. On the other hand, if the ith sector is a small part of the economy,

$$\sum_h \lambda_{hi}$$

would add to less than unity. This suffices to establish the existence of at least one q_{ii} greater than unity. To see this, let $J = \{j \mid q_{ij} \geq 0\}$. Then

$$1 = \sum_j q_{ij}\lambda_{ji} \leq \sum_{j\in J} q_{ij}\lambda_{ji} \leq \max_{j\in J} \{q_{ij}\} \cdot \sum_{j\in J} \lambda_{ji} \leq \max_{j\in J} \{q_{ij}\} \sum_j \lambda_{ji}.$$

If

$$\sum_j \lambda_{ji}$$

is strictly less than unity (ith sector "unimportant" enough), some q_{ij}, at least the

$$\max_{j\in J} \{q_{ij}\},$$

must exceed unity.

The ith row of Q shows what happens to output in the ith sector as each factor endowment rises in turn (and commodity prices are held constant). In general each industry has a "natural enemy" in the sense that for each industry, i, there exists some factor, j, such that industry i would have to contract if only factor j's supply expanded (at constant commodity prices). Of course this is the phenomenon at issue in the Rybczynski result—the actual decline in output as some endowment rises. If the ith sector is relatively unimportant, we have established above that there will be at least one factor which, if it grows, would cause an even greater relative expansion in the ith sector at constant commodity prices.

The columns of Q refer to the impact on all outputs of an increase in a particular V_i (at constant p). If $r > n$, nothing can be said about the columns. For example, consider again the specific-factors model. An increase in the endowment of mobile labor would, at constant p, serve to increase all outputs but perhaps none of them by as great a relative extent as labor.

Although Q is, in the sense described above, less structured than N if the number of factors exceeds the number of commodities, this asymmetry disappears if $n = r$. The reason is that now factor prices are constant if commodity prices do not change. The M-submatrix reduces to an $r \times r$ matrix of zeros. This helps add structure both to the Stolper-Samuelson N-matrix (as we proved above) and the Rybczynski Q-matrix, but primarily to the latter.

From (A20) it is clear that, if M is zero,

$$\lambda Q = I.$$

Since the row sums of λ all add to unity, the type of argument used above serves now to verify that in the even case every column of Q has at least one negative element and, subject to the type of conditions established earlier for N, one element greater than unity. The main body of the paper stated this result somewhat differently: with factor prices uniquely linked to commodity prices and commodity prices held constant, if only one endowment rises, some output must rise by at least a greater relative amount, and some other sector's output must decline.

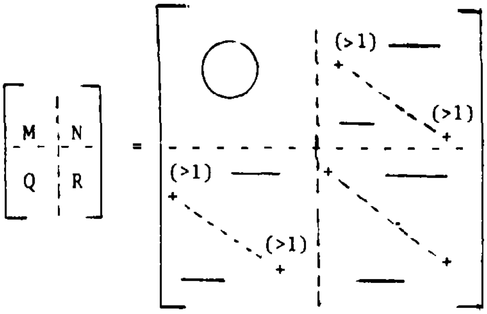


FIG. 4

To conclude, it may prove useful to present two examples of sign patterns of the inverse matrix,

$$\left[\begin{array}{c|c} M & N \\ \hline Q & R \end{array}\right].$$

In the first example, let the number of commodities and factors be the same. Furthermore, impose a strong symmetry on the production structure by assuming that each sector j spends half its revenue on factor j and the other half evenly split among all remaining factors. The sheer symmetry of the case is sufficient to establish the sign and size pattern shown in figure 4. All the features of the two-sector sign pattern would be writ large.

As a second example, consider the $n + 1$ factor, n-commodity version of the specific-factors model, shown in figure 5.[26] The M-matrix has rank 1. The diagonal must be negative—the return to factor i is depressed by an increase in V_i with p held constant. The negative elements off the diagonal (except the last row and column) show that, as the endowment of any specific factor rises (at constant p), the return to *all* specific factors must fall. However, the return to the mobile factor (labor) rises with every positive $\hat{V}_i$ ($i \neq$ labor). This is shown by the first n elements of the last row of M. The last column of M (off-diagonal elements only) shows that an increase in the endowment of mobile labor at constant p must raise the returns to all specific factors.

The R-matrix in this case is especially simple and displays strong gross substitute relationships for outputs along the transformation surface. A rise in p_j increases x_j and reduces all other outputs. (All sectors lose labor to the jth sector.)

The N-matrix shows the Stolper-Samuelson elements. The general result (following from the lack of joint production) is verified: every column of N has at least one element greater than unity (in this case only one) and at least one element negative. The jth column of N shows that, as p_j rises, the return to the factor used specifically in j rises by a greater proportional amount, and the returns to all other specific factors must fall. The positive fractions in the $(n + 1)$st row of N show that each rise in p_j raises the wage rate, but relatively not as much. In the even $(r = n)$ case, each *row* of N would have to contain at least one negative element. In the specific-factors model, the $(n + 1)$st row contains none.

The Rybczynski Q-matrix shows, in the first n columns, that an increase in the endowment of each specific factor would raise output in the sector using that

---

[26] This version was described in Jones (1975).

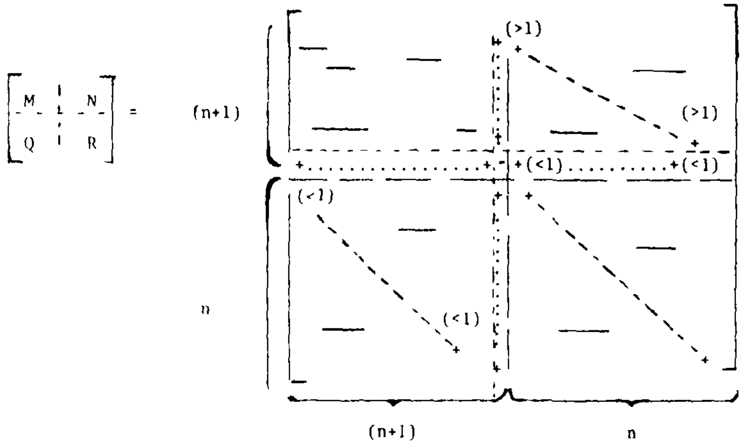


FIG. 5

factor and reduce outputs in all other sectors. However, no element is greater than unity. If V_i should rise by 10 percent, the *i*th sector will attract some labor. But since the wage has been driven up, the labor force in the *i*th sector will rise by less than 10 percent. Therefore output x_i must also rise by less than 10 percent. The last column consists of positive elements, each showing how a sector's output expands if mobile labor grows. A 10 percent increase in the supply of labor causes all sectors to expand. Some may expand by greater than 10 percent if the elasticity of demand for labor in those sectors is especially high. We confirm the general property that each row of Q must contain at least one negative element. In the even case ($n = r$), each *column* of Q would display at least one negative element and one element exceeding unity. In the specific-factors $(n + 1) \times n$ case, the first n columns have no element greater than unity, and the last column has no negative entries.

References

Caratheodory, C. *Calculus of Variations and Partial Differential Equations of the First Order. II. Calculus of Variations.* San Francisco: Holden Day, 1967.

Caves, Richard, and Jones, R. W. *World Trade and Payments—an Introduction.* Boston: Little, Brown, 1973.

Chipman, John. "Factor Price Equalization and the Stolper-Samuelson Theorem." *Internat. Econ. Rev.* 10 (October 1969): 399–406.

Diewert, W. E., and Woodland, A. D. "Frank Knight's Theorem in Linear Programming Revisited." *Econometrica* 45, no. 2 (March 1977): 375–99.

Ethier, Wilfred. "Some of the Theorems of International Trade with Many Goods and Factors." *J. Internat. Econ.* 4, no. 2 (May 1974): 199–206.

Gale, David, and Nikaido, H. "The Jacobian Matrix and Global Univalence of Mappings." *Math. Ann.* 159 (1965): 81–93.

Hahn, Frank. Review of *General Equilibrium Theory and International Trade,* by Takashi Negishi. *J. Internat. Econ.* 3, no. 3 (August 1973): 297–98.

Inada, Ken-ichi. "The Production Coefficient Matrix and the Stolper-Samuelson Condition." *Econometrica* 39 (March 1971): 219-39.

Jones, Ronald W. "Factor Proportions and the Heckscher-Ohlin Theorem." *Rev. Econ. Studies* 24 (October 1956): 1-10.

———. "Duality in International Trade: A Geometric Note." *Canadian J. Econ. and Polit. Sci.* 31 (August 1965): 390 93. (*a*)

———. "The Structure of Simple General Equilibrium Models." *J.P.E.* 73, no. 6 (December 1965): 557-72. (*b*)

———. "A Three-Factor Model in Theory, Trade and History." In *Trade, Balance of Payments and Growth*, edited by Jagdish Bhagwati, Ronald W. Jones, Robert Mundell, and Jaroslav Vanek. Amsterdam: North-Holland, 1971.

———. "Income Distribution and Effective Protection in a Multicommodity Trade Model." *J. Econ. Theory* 11, no. 1 (August 1975): 1-15.

Kemp, Murray. Review of *International Trade*, by Ivor Pearce. *J. Internat. Econ.* 1, no. 2 (May 1971): 251-54.

Kemp, Murray, and Wegge, L. "On the Relation between Commodity Prices and Factor Rewards." *Internat. Econ. Rev.* 10 (October 1969): 407-13.

Lancaster, Kevin. *Consumer Demand: A New Approach.* New York: Columbia Univ. Press, 1971.

Mayer, Wolfgang. "Short-Run and Long-Run Equilibrium for a Small Open Economy." *J.P.E.* 82, no. 5 (September/October 1974): 955-67.

Mussa, Michael. "Tariffs and the Distribution of Income: The Importance of Factor Specificity, Substitutability, and Intensity in the Short and Long Run." *J.P.E.* 82, no. 6 (November/December 1974): 1191-1203.

Pearce, Ivor. *International Trade.* New York: Norton, 1970.

Ruffin, Roy, and Jones, R. W. "Protection and Real Wages: The Neo-classical Ambiguity." *J. Econ. Theory* 14 (April 1977): 337-48.

Rybczynski, T. "Factor Endowment and Relative Commodity Prices." *Economica*, n.s. 22 (November 1955): 336-41.

Samuelson, Paul A. *The Foundations of Economic Analysis.* Cambridge, Mass.: Harvard Univ. Press, 1947.

———. "Prices of Factors and Goods in General Equilibrium." *Rev. Econ. Studies* 21, no. 1 (1953): 1-20.

———. "Ohlin Was Right." *Swedish J. Econ.* 39, no. 2 (December 1971): 365-84.

Uekawa, Yasuo. "Generalization of the Stolper-Samuelson Theorem." *Econometrica* 39, no. 2 (March 1971): 197 217.

Trade Using One Commodity as a Means of Payment

Lloyd Shapley

Rand Corporation

Martin Shubik

Yale University

A general model of noncooperative trading equilibrium is described in which prices depend in a natural way on the buying and selling decisions of the traders, avoiding the classical assumption that individuals must regard prices as fixed. The key to the approach is the use of a single, specified commodity as "cash," which may or may not have intrinsic value. The model, in several variants, is treated as a noncooperative game, in the spirit of Nash and Cournot. The *rules* of the game, including the price-forming mechanism, are independent of behavioral or equilibrium assumptions, which enter, instead, through the *solutions* of the game.

I. Introduction

In the development of a theory of money and financial institutions, there are three kinds of questions which should be clearly differentiated: (1) Why does money come into use, and how? (2) What keeps it in use? (3) What strategic limitations does its use impose on trade, what additional possibilities does it open up, and what are the institutional implications?

The first would involve both historical research and a study of transactions costs and markets. It would be interesting to explore such questions

---

We are grateful for suggestions and constructive criticisms from Pradeep Dubey, Robert Slighton, Rodney Smith, Ross Starr, and an anonymous referee as well as many others with whom we have discussed this model during the past 2 or 3 years. We are also grateful to the National Science Foundation (grant GS-31253) and the Office of Naval Research (contract N00014-76-C-0085) for their support

[*Journal of Political Economy*, 1977, vol. 85, no 5]

© 1977 by The University of Chicago. All rights reserved

as the smallest size or level of trading activity at which a developing economic community should be expected to adopt a commodity money system, a fiat money system, or a banking system, etc. At least since the time of Cournot (1838), it has been observed that, if m commodities are to be bartered on a one-to-one basis with no use of a monetary medium, as many as $m(m - 1)/2$ pairwise trading markets might be needed, while the use of a generally accepted medium of exchange enables the system to function efficiently with only m or $m - 1$ trading posts. Similarly, a money can enable an economic system to function efficiently through purely bilateral transactions, without requiring tortuous sequences of middlemen who pass the goods along to their ultimate consumers (Ostroy and Starr 1974). In other words, a medium of exchange can "de-couple" cross-interactions both between traders and between commodities; and, if an economic society grows large, the social advantages of this soon outweigh any social disadvantages or costs that may be involved. This is of course not the whole story: information costs, search costs, transportation costs, the desire for safety or anonymity, and myriad other details of transactions technology may be invoked as further explanations of why money and financial institutions come into being (see Jones 1976).

Questions concerning the beginnings of money and finance are basically different, however, from the question of why money stays in use. The fish gets on the hook one way and off it in a radically different manner. Once enough people accept a money and a body of law and custom concerning its use, it is difficult, short of social breakdown, hyperinflation, or revolution, for an individual or small group to "opt out" of the system, even if the conditions attending the system's beginnings have changed. The acceptance of the system has a flavor of a noncooperative equilibrium like the "Prisoner's Dilemma": although a person may view, say, fiat money as being of dubious value as a store of wealth, he knows that most others will continue to use it for trade, and he may be in no position to do otherwise himself.

No modern society is a 100 percent monetized society. Segments with barter trade always exist, and the volume of such trade may vary with many factors such as technology, taxes, faith in financial institutions, constraining laws and regulations, and so forth. Nevertheless, if we accept the facts that money is the major means of exchange in a modern economy and that much of the exchange of money for goods, services, and financial instruments takes place via some form of organized market, be it a stockmarket or a supermarket, then we can ask yet another kind of question. What are the restrictions imposed upon economic behavior if we *require* that trade be monetized? How do these restrictions influence the decisions of rational individuals, and how do they limit the range of feasible outcomes? The body of this paper is concerned largely with questions of this type.

Our exposition is based on a multicommodity model of trade as a game in strategic form, to which the noncooperative solution concept of Nash (1951) may be applied. The model is strategically closed, in that, unlike the classical equilibrium models, prices are determined by the actions of the traders, and the system as a whole responds meaningfully to the traders' decisions, even away from equilibrium.[1] The mathematical proof of the existence of a Nash equilibrium and its relationship to the competitive equilibrium of Walras will be presented elsewhere; here we concentrate on an elementary but, we hope, informative exposition, making much use of "Edgeworth" box diagrams. These will enable us to visualize the qualitative effects of monetized trade on such things as price formation, feasibility, Pareto optimality, and equilibrium.

It is worth emphasizing that we are presenting here not a single model with a fixed viewpoint on the nature of money but, rather, a general modeling approach, or framework, within which many different properties and users of money and financial institutions can be analyzed and compared. Our "money" can be a valuable, consumable commodity, or it can be merely paper, of no intrinsic worth. Even in a one-period model, various forms of credit can be introduced, and the trade-inhibiting effects of a limited money supply can be represented. In multiperiod models (not treated here, but see Sec. VIA below), one can go further and introduce lending institutions and derive interest rates. While we make no pretense that we are able to capture all of the complex factors associated with the use of money and financial institutions, it seems reasonable to hope that our approach, extended or elaborated in one direction or another, will prove useful as a basis for more extensive investigations of many facets of this subject.

II. Some Preliminaries on Economics and Modeling

Before proceeding to the description of a specific trading game, it may be useful to discuss some of the highlights of our approach in general terms. In particular, we would like to comment on (*a*) the differences between commodities, commodity money, and fiat money; (*b*) the institutional implications of cash and credit; and (*c*) the distinction between the (legal) rules of the game and the (behavioral) rules of play and the reasons for our selection of the Nash noncooperative equilibrium for the analysis of the latter. In Section VI some other general points will be touched upon, in connection with possible variations and extensions of our model.

---

[1] In his important early paper, Debreu (1952) represents the Walras exchange model as a game in strategic form for the technical purpose of applying a general existence theorem. But as a *descriptive* model his game shares the defect of the Walrasian model of being ill defined, or unrealistically defined, away from equilibrium. Indeed, if only one agent departs from equilibrium, he is presumed to be able to buy and sell at the stated prices, announced by an added fictive player whose objective is to minimize excess demand. But there is no explanation of how the excess demand thereby created is to be satisfied—unless it is out of the bottomless warehouses of the fictive player. See also Arrow and Debreu (1954) and Arrow and Hahn (1971, chap. 5).

A. Money and Commodities

In this paper we are frankly concentrating on the "means of payment" role of money. We cannot contend that this is money's only or even its most important reason for being. But we do argue that this is its most conspicuous function in everyday economic life and that the payment *process* has hardly received a fair share of attention in modern mathematical economics. We shall therefore treat money as an element of strategy (in the game-theory sense) and not merely regard it as either an insubstantial price-reporting or book-balancing abstraction, on the one hand, or just another commodity to be traded and consumed, on the other.

Several authors, in the study of *cooperative*-game solutions to economic models, have had recourse to utility functions of the following special form:

$$U^i(x) = u^i(x_1^i, \ldots, x_m^i) + \lambda^i x_{m+1}^i,$$

where the $(m + 1)$th commodity may be regarded as a kind of transferable utility, or "u-money" (see, e.g., Shapley and Shubik 1966, pp. 807–8; 1969*a*, 1972, 1976; Telser 1972, pp. 4–11; and Aumann and Shapley 1974, p. 180). Although the vague phrase "market game with money" would seem to cover both this approach and our present *noncooperative* model, the connection is only superficial. In cooperative solutions such as the core and the value, the underlying strategic form of the game is irrelevant, and all attention is directed to the actions and potential actions of coalitions. The "u-money," if available, serves only as a vehicle for side payments that adjust the final distribution of utility among traders in coalition. There has been little interest in formulating *cooperative* trading games in which money plays a distinguished role in the trading technology; in fact, core theory (with good reason) has traditionally stood aloof from the processes of trade and price formation.[2]

In our present work, we do not assume that the utility of the payment commodity is additively separable, as above. But it is possible that such an assumption might simplify some of our results or proofs or ensure uniqueness or other good behavior on the part of the noncooperative equilibria, just as it does for the classical competitive equilibria.[3] At least the question merits further study.

---

[2] Telser (1972, chap. 3), however, has analyzed a Cournot-type oligopoly model in terms of the core; see also Shapley and Shubik (1972) for a model whose core has a direct interpretation in terms of prices.

Viewed as a commodity, real or fictitious, money is distinguished in practice by its near-universal acceptability in exchange for other commodities.[4] The reasons for its acceptance lie as much in the realms of the laws and customs of society as in pure individualistic economic reasoning. In an economic model that is not intended to encompass social and legal developments, it may be taken as axiomatic that the monetary good will be accepted at face value according to the existing conventions of the marketplace, regardless of its intrinsic worth or lack of worth. In our present model, we shall *require* that all exchanges be for money. An interesting consequence of this rule is that the set of attainable redistributions of goods will fail to include many redistributions that would be possible if arbitrary, transaction-cost-free barter were allowed.[5] It should be noted that this curtailment also occurs under the classical rules of Walrasian exchange at stated prices. This is not necessarily unrealistic. A similar curtailment of the feasible set occurs in practice, we believe, in most societies that rely on organized markets and relatively stable price systems for the redistribution of goods.

Because of its general acceptability a means of payment has obvious value to its holder, but this value need not show itself in his utility function.[6] Indeed a *fiat money* would enter the utility function only through the effect of truncation: unspent money at the game's end may be presumed to have buying power in the world to come. (You *can* take it with you!) A *commodity money*, on the other hand, like gold or silver, has utility in its own right, to which we may add any extra buying power it may be considered to possess—an alternative or incremental value conferred by society's acceptance of its special monetary role.[7]

It is obvious that an adequate investigation of the utility of money demands a dynamic treatment in a multiperiod model, with the possibility of durable as well as perishable goods. We forgo any such investigation in this paper, adopting, instead, the expedient of reserving a place in the utility functions for the "means of payment" without insisting that it actually be consumable or have utility. Within this framework we can encompass the range between a commodity money of great intrinsic worth, at one extreme, and a perishable fiat money, at the other.

---

[3] With an additively separable "u-money" in sufficient supply, the competitive allocations of the other goods are just those that maximize the sum $\sum u^i(x)/\lambda^i$; thus a "fixed-point" situation reduces to a simpler maximization problem.

[4] Clower (1967) goes so far as to assert, as a matter of definition, that *any* commodity universally acceptable in exchange is a "money."

[5] See the end of Sec. IIC.

[6] Many economic investigations include a "money" in the utility functions; see, e.g., Patinkin (1956), Telser (1972 and elsewhere), or the discussion in Samuelson (1947, pp. 119-21).

[7] Such an adjustment for buying power may be called for even in the case of pure barter. A person's utility for a large quantity of wheat, for example, evaluated at a particular point in time, may depend in part on his expectation of being able to trade it later for something more directly useful to him. It is difficult to treat this subject with any precision in a static, one-period model.

B. Cash and Credit

Our point of departure will be a basic model where "cash" payments are required in advance on all purchases. We then explore some ways of relaxing this condition, via a sort of shopkeepers' credit in the form of deferred payments secured by expected receipts. Since only a one-period model is considered, interest rates and the money market cannot be fairly represented. Nevertheless, some institution-modeling problems do appear, since beyond a narrow zone of "conservative" credit (see Sec. IVB below) the possibility of insolvency and default exists and must be faced if the game is to be well defined. It may seem overelaborate to burden the abstract model with the details of a bankruptcy proceeding in which the assets of the trader who cannot pay his bills are liquidated to satisfy the creditors. Yet important properties of the game and its equilibria may well depend on just how such details are handled, once we relax the "cash on the barrelhead" rule.[8] One simple expedient is to allow negative holdings of "cash" at the game's end while postulating a disutility (suitably concave and continuous) for being "in the red." Implicitly there are loan sharks, swimming outside the model, who pay off the creditors while making life uncomfortable for the debtors, but the burden of providing detailed rules for insolvency is avoided.

Even without default there are institutional overtones in the granting of credit. Suppose a trader pays for some goods with promissory notes rather than cash, expecting to redeem them with cash received from the sale of his own goods. What if some of his customers also use promissory notes? When and how does the redistribution of the actual "means of payment" commodity take place? It appears that a central clearinghouse must be created. This would represent a major change in the nature of the model, which—as will be seen—is otherwise quite clearly decentralized, with respect to both traders and commodities.

While we cannot enter into a detailed discussion here, it is important to note also that credit generally involves a *contract* between two parties, whereas cash generally does not (except perhaps in the weak sense that an individual holder of a dollar bill may regard himself as a creditor of the government). We may for convenience describe a person's indebtedness as "negative cash," but it is basically a different instrument. In a society where all pay cash, default and the laws and procedures for dealing with it need not be considered. A society with even the simplest forms of credit is fundamentally more complex than one without credit, and it can adequately be represented only by a fundamentally more complex model.

---

[8] The rule proposed by Postlewaite and Schmeidler (1975), which in effect confiscates and destroys all goods belonging to the bankrupt party, seems both drastic and unrealistic (particularly away from equilibrium), since avoiding insolvency in that model depends not only on personal prudence but on correctly estimating the actions of the other traders.

C. *Rules and Solutions*

The game theorist is usually at great pains to assure himself that the rules of the game are completely defined before he turns to the problem of solving it. The reason for this caution stems from the fact that, while the *descriptive theory*—covering the moves and strategies and information and payoffs, etc.—can be "hard" and mathematically precise, the solution theory is often "soft" and indeterminate, since it expresses the actions of sophisticated, free-willed decision makers. In the multiperson non-zero-sum games common in economics, this indeterminacy shows itself both in the multitude of different solution *concepts* that game theory offers for consideration—like core, value, bargaining set, noncooperative equilibrium, etc.— and in the nonuniqueness of outcome that is so often exhibited by the actual solutions, under any one of these concepts.

We would like to stress, therefore, that we are presenting a well-defined game in the descriptive sense, formulated independently of any assumptions of equilibrium or of what might or might not be "rational" behavior. When the players have made their individual decisions, the market prices and the transfers of goods are *completely determined.* To this game a variety of solution concepts might be applied, but the rules of the game and its solution are two different things.

Having said this, we must admit that we have built our model in a particular way (in a "strategic" form, as opposed, say, to a "coalitional" form) in anticipation of applying a particular solution concept—the "Nash" noncooperative equilibrium. In much of our previous work in this general area we have, instead, used the coalitional form and (with many others) have considered static, essentially noninstitutional games of exchange and/or production, establishing some interesting links between the competitive equilibrium and the core or other *cooperative* game concepts.[9] However, this kind of coalitional approach fails to capture the dynamics of price formation and the essentially individualistic nature of much economic decision making, and so further investigations using *noncooperative* game concepts now seem to be in order.

---

[9] See, e.g., Shapley and Shubik (1966, 1969*a*, 1969*b*, 1976) or the books of Arrow and Hahn (1971), Scarf (1973), Aumann and Shapley (1974), and Hildenbrand (1974), where many additional references will be found.

The classical model is content to take prices as though given by an "invisible hand," insensitive to the actions of the traders at least in the short run. There is an implicit understanding, not reflected in any mathematical assumption of the classical model, that the traders are so numerous and their individual resources so small that this insensitivity is a good approximation to reality. Our model, on the other hand, has prices that depend in a reasonable way on the individual trading decisions. They are driven upward by increased buying and downward by increased selling. This puts us into a position to examine the validity of the classical assumption of unyielding prices when the traders are *not* individually insignificant and, indeed, to explore the transition zone between perfect competition à la Walras and oligopolistic competition à la Cournot. But to carry out this program, our game-theoretic solution concept must be strategy oriented and collusion free, and the Nash noncooperative equilibrium is ideally suited to this purpose.

It may be of interest to point out in conclusion that if we were to reduce our present model to its characteristic function,[10] we would get *a different cooperative game* from the unrestricted-barter or "Edgeworth" game that is usually considered in connection with the core. The reason is that our rules cause all trade to use a single set of prices. If goods, or goods and money, pass between Traders 1 and 2 in a certain ratio, then they cannot pass between 1 and 3 in a different ratio. From the coalitional standpoint, this restriction on trade can lead to a paradoxical result—a failure of superadditivity. In fact, if we make the natural assumption that subeconomies can form and establish their own price systems, then an economy that is fractured into opposing coalitions may be able to reach allocations on the Pareto surface that cannot be achieved by the same economy united.[11]

III. The Basic Model

In order to describe a well-defined game of exchange in which a specific commodity is used as a means of payment, we must spell out how the prices are formed. The classic general equilibrium model is content to establish the existence of prices (often not unique) *at equilibrium*. For a proper game model, however, we need rules that determine the prices for positions of disequilibrium as well. With every player free to make an independent decision, the model must yield a well-defined output for every set of inputs.

---

[10] The "characteristic function" of a game characterizes the outcomes that can be achieved by any subset of players. See von Neumann (1928), Aumann and Peleg (1960), Scarf (1967), Shapley and Shubik (1973), etc.

[11] It is not generally realized that there is a distinctive "Walras" cooperative game, with a more restrictive characteristic function than the Edgeworth game. Despite the possible failure of superadditivity, and hence of balancedness, the existence of the core in the Walras game is not threatened, since the competitive equilibrium must still have the core property. But the Walras core and Walras-Pareto set will in general be different from the Edgeworth core and Edgeworth-Pareto set. Superadditivity can be restored to the Walras game by the device of taking the "superadditive cover" (see Shapley and Shubik 1969*a*), but the result is still not the Edgeworth game.

Several types of price-forming mechanisms might be considered, each placing different restrictions on the strategic possibilities. In particular the traders might control only the quantities offered or demanded, or they might name reservation prices or price ranges or even complete demand curves for their individual transactions. Multistage bargaining might be introduced, or a centralized procedure that converts a set of unilateral price declarations into a unique, market-wide price for each good.[12] Here, as we are interested in general, anonymous exchange, with market-wide prices but with a minimum of ad hoc institutional detail, we adopt what is essentially a generalization of Cournot's original approach. The strategic variables will be *quantities*, not prices, but they will include quantities of the special good that serves as "cash" or "trading money." Indeed, in our simplest version, *only* that good will be subject to strategic choice.

Formally, in our prototype model we shall assume that there are n traders trading in $m + 1$ goods, where the $(m + 1)$th good has a special operational role in addition to its possible utility in consumption. We attribute to each trader an initial bundle of goods,[13]

$$a^i = (a^i_1, \ldots, a^i_m, a^i_{m+1}),$$

and a concave utility function,

$$u^i(x^i_1, \ldots, x^i_m, x^i_{m+1}).$$

We emphasize that u^i need not actually depend on x^i_{m+1}; the possibility of a fiat money is not excluded.

The general procedure will be for the traders to put up quantities of the first m goods to be sold and simultaneously to put up quantities of the $(m + 1)$th good to buy them, all at prices determined by the market-wide supply and demand for each good. For expository purposes, our prototype model will require the traders to offer for sale *all* of their holdings of the first m goods, though they need not spend all of their $(m + 1)$th good. A trader may well buy some of his original goods back, but they must go through the market. In other words, in this version of the model the trader does not own his initial bundle outright; he merely owns a *claim on the proceeds* when the bundle is sold.[14]

---

[12] Levitan and Shubik (1971, 1972) have given examples of different strategy spaces and Pareto surfaces that are possible in oligopolistic models.

[13] Superscripts will be used consistently to denote traders. Summation over traders will be denoted by a horizontal bar; thus $\bar{a}_j$ means

$$\sum_{i=1}^{n} a^i_j.$$

[14] This simplifying condition is not as unreasonable as it might appear at first glance. In a multiperiod context, it would amount to requiring that all "paper" profits or losses be realized every trading period. Some other modeling possibilities will be discussed in Sec. VI below.

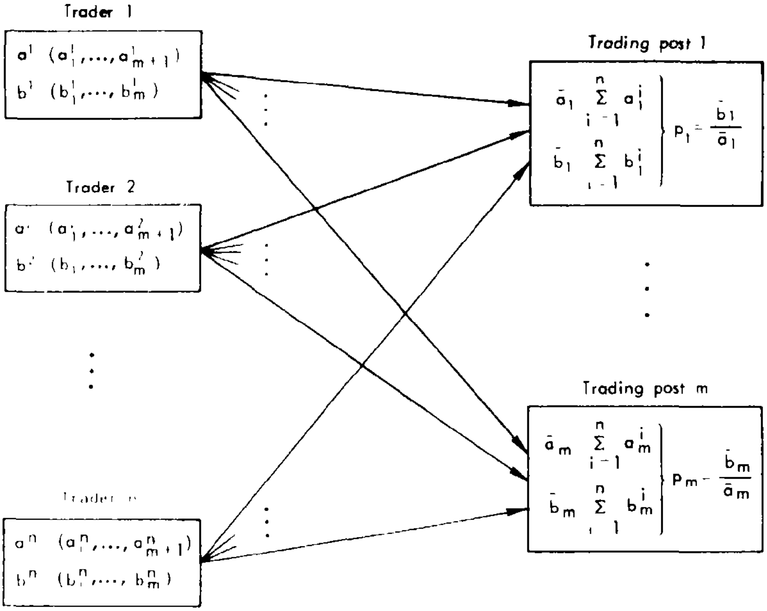

Trader 1
$a^1\ (a^1_1,\ldots,a^1_{m+1})$
$b^1\ (b^1_1,\ldots,b^1_m)$
Trader 2
$a^2\ (a^2_1,\ldots,a^2_{m+1})$
$b^2\ (b^2_1,\ldots,b^2_m)$
Trader n
$a^n\ (a^n_1,\ldots,a^n_{m+1})$
$b^n\ (b^n_1,\ldots,b^n_m)$
Trading post 1
$\bar{a}_1 = \sum_{i=1}^{n} a^i_1$
$\bar{b}_1 = \sum_{i=1}^{n} b^i_1$
$p_1 = \frac{\bar{b}_1}{\bar{a}_1}$
Trading post m
$\bar{a}_m = \sum_{i=1}^{n} a^i_m$
$\bar{b}_m = \sum_{i=1}^{n} b^i_m$
$p_m = \frac{\bar{b}_m}{\bar{a}_m}$

FIG. 1.—Decentralized trade

Let us imagine m separate *trading posts*, one for each of the first m commodities, where the total supplies $(\bar{a}_1, \ldots, \bar{a}_m)$, assumed to be positive, have been deposited for sale "on consignment." Figure 1 illustrates this. Each trader i makes bids by allocating amounts b^i_j of his $(m+1)$th commodity among the m trading posts, $j = 1, \ldots, m$. We shall denote his *strategy*, in the game-theoretic sense, by the vector $b^i = (b^i_1, \ldots, b^i_m)$. There are a number of possible rules governing the permitted range of bids. In the simplest case, with no credit of any kind, the limits on b^i are given by

$$\sum_{j=1}^{m} b^i_j \leq a^i_{m+1},$$

and

$$b^i_j \geq 0, \qquad j = 1, \ldots, m.$$

The interpretation of this spending limit is that the traders are required to pay *cash in advance*. More generally, we might allow them to defer payment, either in anticipation of receipts or under some other credit arrangement that would have to be made explicit in the model.[15]

---

[15] See Sec. IVB below. The simplest way to model credit is to remove the spending limit entirely, while extending the domain of the utility function to include negative

A. Price Formation

The prices now emerge in a natural way as a result of the simultaneous bids of all buyers; we define

$$p_j = \bar{b}_j/\bar{a}_j, \qquad j = 1, \ldots, m.$$

Thus bids precede prices. Traders allocate their budgets *fiscally*, committing quantities of their means of payment to the purchase of each good without definite knowledge of what the per-unit price will be. At an equilibrium this will not matter, as prices will be what the traders expect them to be. In a multiperiod multitrader context, moreover, the traders will know the previous prices and may expect that fluctuations in individual behavior in a mass market will not change prices by much. But any deviation from expectations will result in changing the quantity of goods received rather than the quantity of cash spent. In practice, if one allocates a portion of one's budget for purchase of a certain good in a mass market, this will be different—but not too different -from a decision to buy a specific amount at an unspecified price. It is a matter of letting one's stomach rather than one's purse absorb the fluctuations.

The prices in our model are so determined that they will exactly balance the books at each trading post. Note that there is no need for a central clearinghouse and there are no interpost or intertrader transactions. The amount of the jth good that the ith trader receives in return for his bid b_j^i is

$$x_j^i = \begin{cases} b_j^i/p_j & \text{if } p_j > 0, \\ 0 & \text{if}^{16}\ p_j = 0. \end{cases}$$

His final amount of the $(m + 1)$th good, taking account of his sales as well as his purchases, is

$$x_{m+1}^i = a_{m+1}^i - \sum_{j=1}^{m} b_j^i + \sum_{j=1}^{m} a_j^i p_j.$$

His *payoff*, in the game-theory sense, must be expressed as a function of all the traders' strategies; accordingly, we write

$$\Pi^i(b^1, \ldots, b^i, \ldots, b^n) = u^i(x_1^i, \ldots, x_{m+1}^i).$$

---

amounts of the $(m + 1)$th commodity so as to provide a suitable *disutility* for being caught short of cash when the sales and purchases are all added up. To close the model, an outside source of cash would have to be postulated, to cover the payments due the other traders when one trader defaults. In a dynamic, multiperiod context, we can conceive of durable goods, other than the $(m + 1)$th "payment" commodity, that are carried forward and used as security for the granting of credit for commercial loans (Shubik 1977).

[16] Note that $p_j = 0$ implies $b_j^i = 0$. Thus a trader receives nothing if and only if he bids nothing.

Here the x's depend on the b's according to the three preceding displayed equations. It is noteworthy that the function Π^i is *concave* in b^i for each i; this is important for the existence proof.

Because of the mechanism for price formation and the anonymous allocation of sales, all traders pay the same price for the same good. However, the operation of the market system gives rise to what have been called *pecuniary externalities*, in the sense that the prices paid by one trader are dependent on the monetary actions of the others (Viner 1931; Shubik 1971). In the classic barter market this is not so; trading pairs or groups can form and exchange goods unaffected by the actions of others.

B. The Noncooperative Solution

A noncooperative or "best-response" equilibrium has been defined in game theory as a set of strategy choices by the players with the property that no player, given the choices of the others as fixed, can gain by changing his own choice.[17] Specialized to the present application, this solution concept will be recognized by economists as a close relative of the classic Cournot oligopoly solution (cf. Mañas 1972). It will consist of an n-tuple of strategies:

$$\hat{b} = (\hat{b}^1, \hat{b}^2, \dots, \hat{b}^n)$$

such that for each i the function $\Pi^i(\hat{b}^1, \dots, b^i, \dots, \hat{b}^n)$, considered as a function of b^i alone, is maximized at $b^i = \hat{b}^i$.

The following existence theorem for the noncooperative equilibrium (NE) may be proved by an application of the Kakutani fixed-point theorem to the "best-response" adjustment process or transformation: $b \rightarrow b'$, with each b'^i being a best response by player i to the set of choices $\{b^j : j \neq i\}$. The details of the proof, which are rather lengthy, will be given elsewhere.

> *Theorem 1.* For each trader, $i = 1, \dots, n$, let u^i be continuous, concave, and nondecreasing. For each good, $j = 1, \dots, m$, let there be at least two traders with positive initial endowments of good $m + 1$ whose utility for good j is strictly increasing. Then a noncooperative equilibrium exists.

Note that there is no assumption that good $m + 1$ has intrinsic value to anyone. It must merely be available to enough people so that nontrivial markets for the other goods can be formed.

---

[17] The definition of Nash (1951), who, however, includes the possibility of mixed strategies. These have no plausible interpretation in our present model, so we shall be searching only for "Nash equilibria in pure strategies." The term "best-response equilibrium" is due to Robert Wilson.

Our reasons for concentrating on the noncooperative equilibrium rather than on one of the cooperative solutions of game theory have already been discussed in Section IIC. We would like merely to reiterate here that at least one purpose of our work is to formulate and study the *descriptive* game model, without reference to any particular solution concept. Indeed much of the discussion that follows concerns what are sometimes called "pre-solution" concepts like feasibility, Pareto optimality, or individual rationality, which do not depend on equilibrium notions at all.

IV. The Edgeworth Box

The case $m = 1$, $n = 2$ lends itself to simple two-dimensional descriptive analysis based on the familiar Edgeworth box. To avoid the confusion of too many lines and curves in one place, we shall make a sequence of diagrams, using the same labels as far as possible. As we shall see, much of this geometry will apply also to the general case, with many goods and traders, because of the way in which the operation of the system "decouples" both the traders and the trading posts.

In figure 2, the first trader's holdings are measured up and to the right of the point O^1, while the second's are measured down and to the left of O^2. The dimensions of the box are $\bar{a}_2$ high by $\bar{a}_1$ wide. The point R represents a typical *initial allocation*; thus we have $a_1^1 = |M^1R|$, $a_2^1 = |G^1R|$, etc.[18] The points S^1 and S^2 represent typical *strategy choices*; thus we have $b_1^1 = |M^1S^1|$ and $b_1^2 = |M^2S^2|$. They are restricted to lie along the edges M^1O^1 and M^2O^2, respectively, since we are requiring all of good 1 to be sent to market. (If we did not make this restriction, the strategies would lie arbitrarily in the rectangles $RM^1O^1G^1$ and $RM^2O^2G^2$, respectively; see Sec. VI below.)

Now consider the line joining S^1 and S^2. Its slope is $(b_1^1 + b_1^2)/(a_1^1 + a_1^2)$, which is just the price p_1. Moreover, it divides the line M^1M^2 into two segments which are equal in length to the amounts x_1^1 and x_1^2 of good 1 purchased by Traders 1 and 2, respectively. Similarly, it divides the line G^1G^2 into segments equal to their final holdings x_2^1, x_2^2 of the monetary commodity, or "cash." We see therefore that the *final allocation* is represented by the point F. The vector RF represents the actual transaction that takes place; its slope, naturally, is equal (in absolute value) to the price p_1.

It is essential that the reader understand the nomogram demonstrated in figure 2, as it is the key to all the diagrams that follow.

Figure 3 shows the effect of holding S^2 fixed while varying S^1. As S^1 moves over the interval from M^1 to O^1, the point F traces out the curve A^1RB^1. This curve is a portion of a hyperbola whose asymptotes are the

---

[18] The notation $|PQ|$ indicates the length of segment PQ.

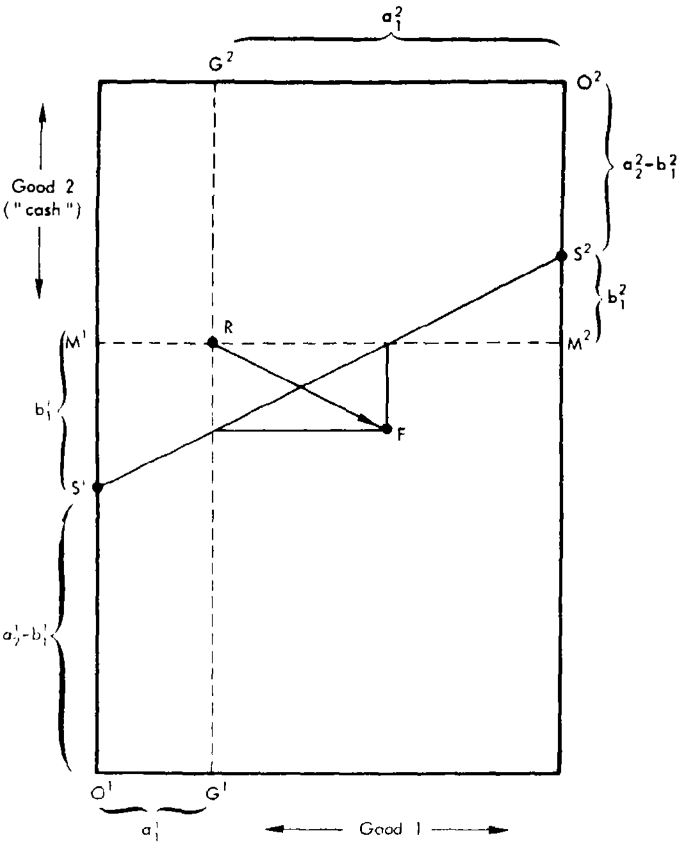


FIG. 2.—Price formation

horizontal and vertical lines through S^2. If we reverse the process and move S^2, holding S^1 fixed, we trace out the curve A^2B^2, which is part of a similar rectangular hyperbola centered at S^1. Endpoints A^1, A^2 correspond to zero bids, while endpoints B^1, B^2 reflect the upper limit on the amount a trader can bid. It happens in this case that we did not allow Trader 2 enough cash to be able to buy back his original holding when Trader 1 plays S^1, so the curve A^2B^2 stops short of R.

These traces are comparable to the price rays or "budget sets" that confront a trader in the classical Walrasian model with its fixed prices. The difference is that in the present case the price is not constant but reacts to variations in a trader's own decisions, so we get a curve instead of a line. The curve is concave, as one would expect, so that if a trader increases his purchase he drives the price up, while if he bids less the price

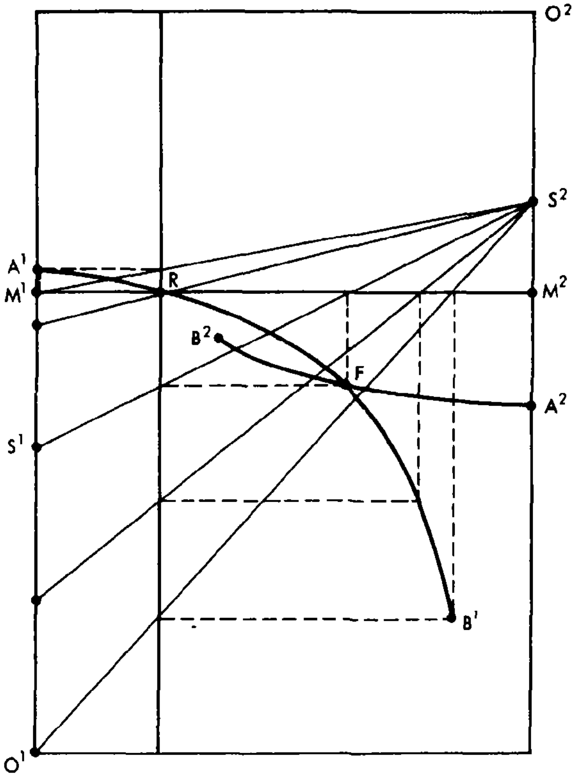


FIG. 3.—"Budget sets"

falls. The connection between the two approaches may be illustrated in figure 3 by supposing the second trader's holdings and bid to be very large, pushing points O^2 and S^2 far off the page. The first trader's hyperbolic "budget set" would then approximate a straight line through R and F, reflecting the fact that he now has little influence on the price.

A. Two Kinds of Equilibrium

So far we have discussed only the mechanics of the rules of exchange. We are now ready to add the traders' preferences to the picture, by superimposing the contours of the utility functions u^1 and u^2. As shown in figure 4, S^1 happens to be the *best response* to S^2, since A^1B^1 is tangent at F to the contour of u^1. (Note that the curvature of A^1B^1 is such that there is always a unique point of tangency.) Similarly, S^2 is the best response to S^1, since A^2B^2 is tangent at F to one of the contours of u^2. Thus figure 4

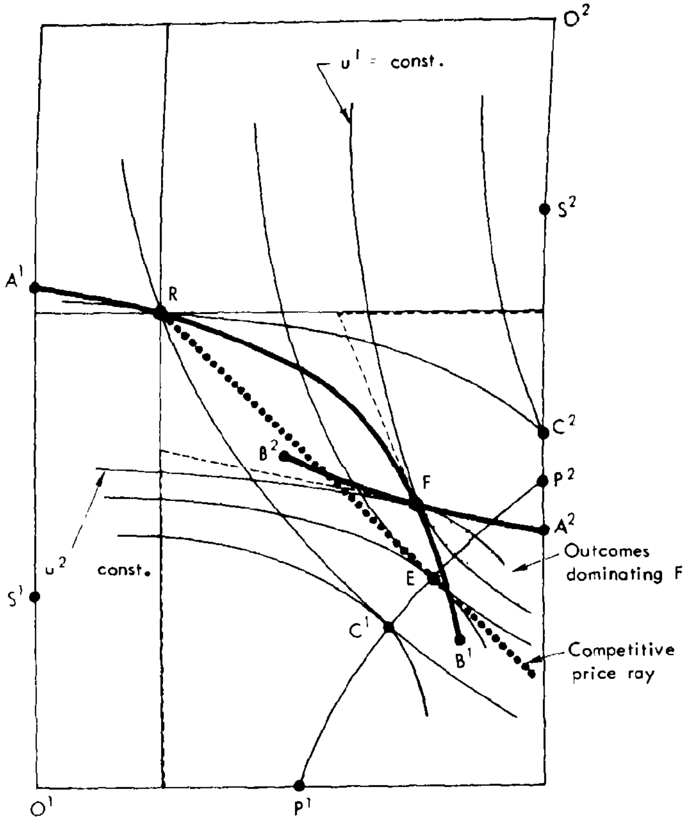


FIG. 4. Noncooperative (F) and competitive (E) equilibrium

illustrates a noncooperative or "Nash" equilibrium (NE) for the market. Neither trader, knowing the strategy of the other, would wish to change.

A striking feature of this kind of equilibrium is its nonoptimality. Since curves A^1B^1 and A^2B^2 are not generally tangent to each other, the point F cannot be expected to be Pareto optimal or "efficient."[19] In effect, the traders are working with unequal *marginal* prices, represented by the unequal slopes of A^1B^1 and A^2B^2 at F. Any outcome in the shaded region would be preferred by both traders to the NE allocation at F. In particular, they would both profit from increased trade at the *average* price p_1, represented by the slope of RF.

The reader familiar with the Edgeworth diagram will recognize the

---

[19] Indeed Pareto optimality in a NE can only occur at corners of the indifference curves or in special cases like $F = R$ or $F = B^2$.

contract curve $C^1EP^2C^2$ in figure 4, which is a subset of the more extensive *Pareto set* $O^1P^1C^1EP^2C^2O^2$. The competitive equilibrium (CE) is represented by the competitive price ray RE, which is tangent to both indifference curves at the competitive allocation E. The situation illustrated seems to be typical: there is less volume of trade at the NE than at the CE. But the reverse is also possible. In fact, by a somewhat contorted but perfectly legitimate arrangement of the indifference contours, we could make any given point outside the dotted lines in figure 4 the location of a unique CE, while keeping a unique NE at F.

Although we have not illustrated it, it is also not difficult to construct utility functions for which several distinct strategy pairs (S_1, S_2) are in noncooperative equilibrium. This may or may not be accompanied by a corresponding multiplicity of CE; the two kinds of solution are not directly interlinked. But their general mathematical properties are quite similar. For example, we would expect that even for general m and n, if the utility functions are smooth and the initial allocation is chosen according to a nonatomic probability distribution, then with probability 1 there will be a finite, odd number of NE (cf. Debreu 1970).

B. The Feasible Set and Credit

The set of feasible outcomes can be determined by holding one trader's bid at its upper extreme O^i or lower extreme M^i and sweeping through the strategies of the other trader. The feasible set is shown in figure 5 (solid shading); for example, outcome L results if both traders bid their upper limit.[20] This feasible set has nothing to do with rationality or motivation; it merely describes those outcomes that the mechanism *can* be made to produce. It necessarily contains all NE allocations, but, as for the CE allocations, we know a priori only that they lie in the quadrants of the Edgeworth box "northwest" or "southeast" of the point R. It may well be that some or all of the CE allocations fall outside of the feasible set, under our present trading rules.

By introducing *credit* we can enlarge the feasible set, permitting the traders to bid more of the "means of payment" commodity than they actually possess as cash on hand. In diagrammatic terms, the first trader, say, could be allowed the segment C^1M^1 instead of O^1M^1 (fig. 5). The

---

[20] Most boundary points are attainable, but those on the line M^1M^2 are not, except for R itself. There is, however, an exceptional "null" outcome, not in the Edgeworth box, in which Trader 1 bids and gets M^1 and Trader 2 bids and gets M^2. Our rule is that if a trader bids nothing he gets nothing, even if the price is zero. So if both traders bid nothing, the entire stock of good 1 is lost. (We may imagine that it goes to an otherwise unnoticed "scavenger," who sends infinitesimal bids to each trading post in the hope of making a killing.) This exceptional outcome is manifestly undesirable and unstable, and it has no real effect on the solutions of the game, though it causes technical difficulties in the existence proof.

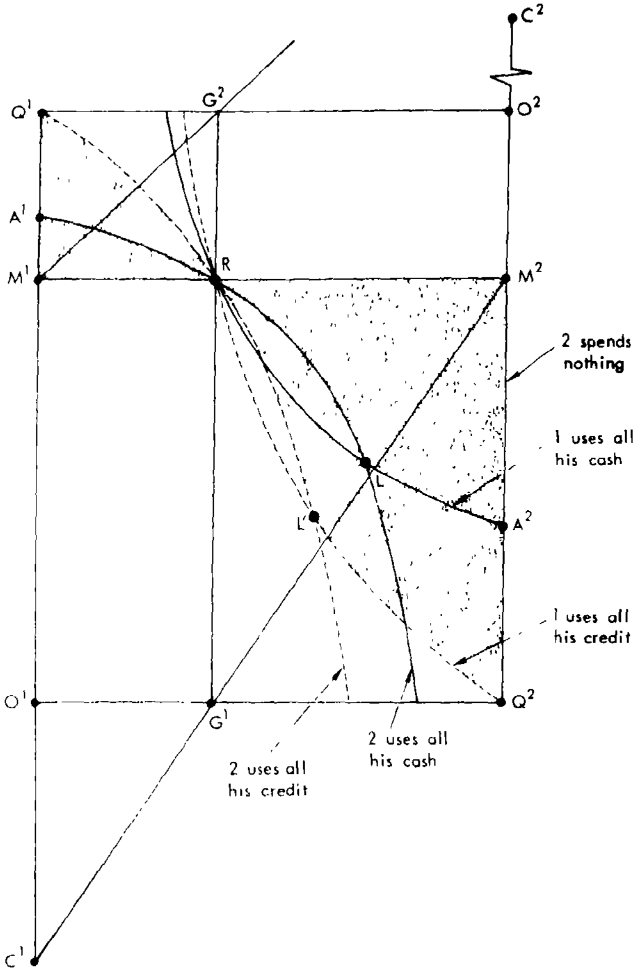


FIG. 5.—The feasible set with and without credit

length $|O^1C^1|$ has a simple interpretation in terms of banking: it is the largest amount of good 2 that could be loaned to Trader 1 with the certainty that he will be able to repay. To see this, consider that the worst case is $S^1 = C^1, S^2 = M^2$, with Trader 1 bidding the limit and Trader 2 bidding nothing. This leads to outcome Q^2, at which Trader 1 is just barely solvent. A similar "conservative" credit limit C^2 for Trader 2 can be determined in the same way.[21]

---

[21] It will become clear (Secs. IVD and VA) that when there are many traders the amount of "conservative" credit is likely to be very small, since it depends on the amount

If we alter the rules of the game to enable both traders to use this conservative banking credit, the feasible set is extended as indicated by the stripes in figure 5. For example, outcome L' results if both traders bid up to their new limits. The new feasible set still does not fill up the two basic quadrants, however, and so we still cannot be sure that any competitive outcomes will be attainable.

With more liberal credit, the feasible set would cover a larger part of the two basic rectangles and eventually all of them, but it would also include areas above and below the Edgeworth box, representing situations in which one or the other trader is "caught short." If we regard credit as the issue of a financial instrument, the interpretation of a point outside the box is simple and familiar—it amounts to stating that after trade an individual ends up with none of the monetary commodity in hand and with outstanding claims that he cannot meet. In order to complete the model we must either describe the utility to any player at such an outcome or add further moves to the game, corresponding to bankruptcy proceedings (see Sec. IIB).

C. *Individual Rationality*

In the present version of our model we are requiring that all of the initial endowments, except for the "payment" commodity, be put up for sale.[22] Consequently the concept of ownership is somewhat different from that usually assumed in microeconomics and is closer to that in law. All goods are monetized, and trade is virtually anonymous. The economy is an accountant's dream, but by forcing goods to pass through the market ownership rights are weakened. In particular, it may be impossible for a person to recover the bundle he originally started with or obtain an equivalent bundle. On the other hand, he may be able unilaterally to guarantee an outcome superior to his initial holding. (This contrasts with the classical Edgeworth and modern "core" models, where an individual trader can always defend his initial holdings if he so desires, but no more.)

The "defensive" possibilities are easily illustrated. In figure 6 the first trader is fortunate, since he has relatively little of good 1 that must go to market and relatively much of good 2 that he can use at his discretion. He cannot, of course, protect his exact initial holding, since this would mean being able to enforce outcome R regardless of the other's choice of S^2. But he can force the outcome to *have as much utility* for him as R. In fact, by playing S^1 as shown, he restricts the possible outcomes to the set A^2B^2, a set which lies entirely on the "high" side of the indifference

---

of a trader's own money that will surely come back to him from his own bids. In this respect our two-trader exposition may be misleading. We may also remark that "conservative" credit does not require the aid of a central clearinghouse (see Sec. IIB), as any promissory notes can be redeemed at each trading post independently.

[22] But see Sec. VI, where this requirement is removed.

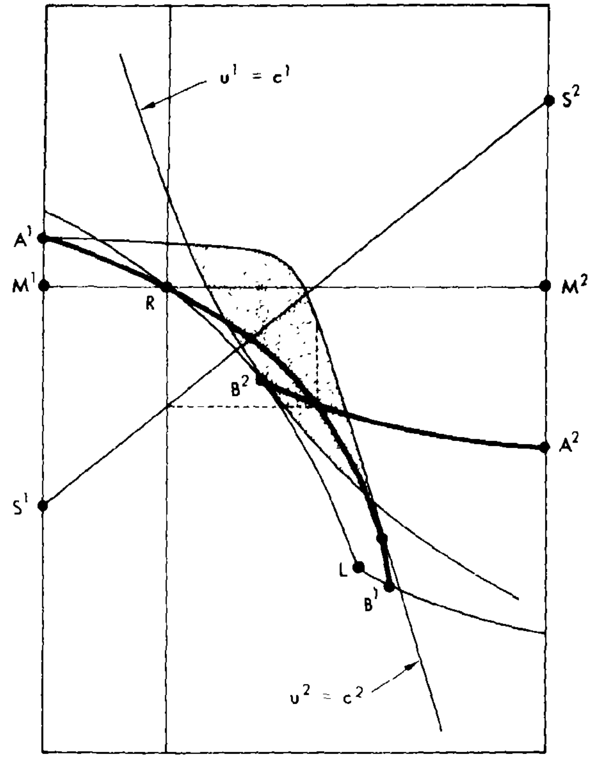


FIG. 6. "Max-min" strategies and the individual rationality zone

curve $u^1 = c^1$, as drawn. This is in fact the best he can do defensively: S^1 is what would be called his "max-min" strategy, and c^1 is his "max-min" payoff. We see that moving S^1 either up or down would move B^2 into a region of lower payoff, either toward R or toward L.

The second trader, on the other hand, has a relatively poor defensive position. The highest u^2 indifference curve that completely contains one of the A^1B^1 sets is the curve $u^2 = c^2$, which is distinctly inferior to R for him. There are two critical points: the endpoint A^1 and the tangency point near B^1. As S^2 moves, the A^1B^1 curve rotates on R. Raising S^2 would make A^1 worse, while lowering S^2 would make the critical point near B^1 worse, so S^2 is the "max-min" strategy.

The shaded area in figure 6 represents the so-called individually rational allocations, bounded by the curves $u^1 = c^1$ and $u^2 = c^2$. The *feasible* individually rational allocations are just those that lie beneath the horizontal line RM^2, and it is a theorem that every NE solution will lie in that region. In contrast, the Walras-Edgeworth individually rational zone, which contains every CE solution, would be defined by the two indifference curves through R (not shown). It is easy to see that this region could be entirely disjoint from the other. This is not too surprising, as the two regions arise from games with the same economic data but different rules of trade.

D. Many Traders

The same diagrams can be used, one-sidedly, when there are more than two traders. This is because our model has the *aggregation* property, with both strategies and outcomes being additive over traders. Viewed by the other traders, a set or coalition of traders acting together is hardly distinguishable from a single, larger trader. In the diagrams, we may regard the point S^2 not as a single bid but as the vector sum of bids by Traders 2, 3, . . . , n, using R as the "origin" from which the vectors are defined. The first trader may notice a quantitative but not a qualitative difference from the two-person case: since the other traders combined may have far more resources than he does alone, the distance of S^2 from R and S^1 may make the slope of S^1S^2 (i.e., the price) almost insensitive to his own choice.

It might appear at first glance that the *feasible set*, when there are many traders, would continue to be a sizable fraction of the set of outcomes that are feasible under unrestricted, Edgeworthian barter. But this is not the case. Let us count dimensions. The unrestricted allocation space has dimension $(m + 1)(n - 1)$.[23] Even in the case $m = 1$, there are $2n - 2$ dimensions of possible outcomes if trade is not required to pass through the "trading post" mechanism. This compares with the n dimensions at most that can arise when n traders bid by each selecting a point on a line as his strategy. Thus, starting at $m = 1$, $n = 3$, the feasible set is only a lower-dimensional surface or manifold in the set of outcomes that would be possible under unrestricted trading.

E. Many Commodities

Since the trading posts operate essentially independently of each other, it is possible to continue to use these diagrams when there are more than two commodities (i.e., more than one trading post). But there are two ways in which the markets remain intercoupled: (1) through the spending limits, which apply to all bids combined, so that a trader's upper bound at each trading post will depend on what he spends at the others, and (2) through the utility functions, which in general reflect complementaries,

---

[23] The "+ 1" here is for the monetary good, while the "− 1" reflects the fact that the sum of the holdings is constant.

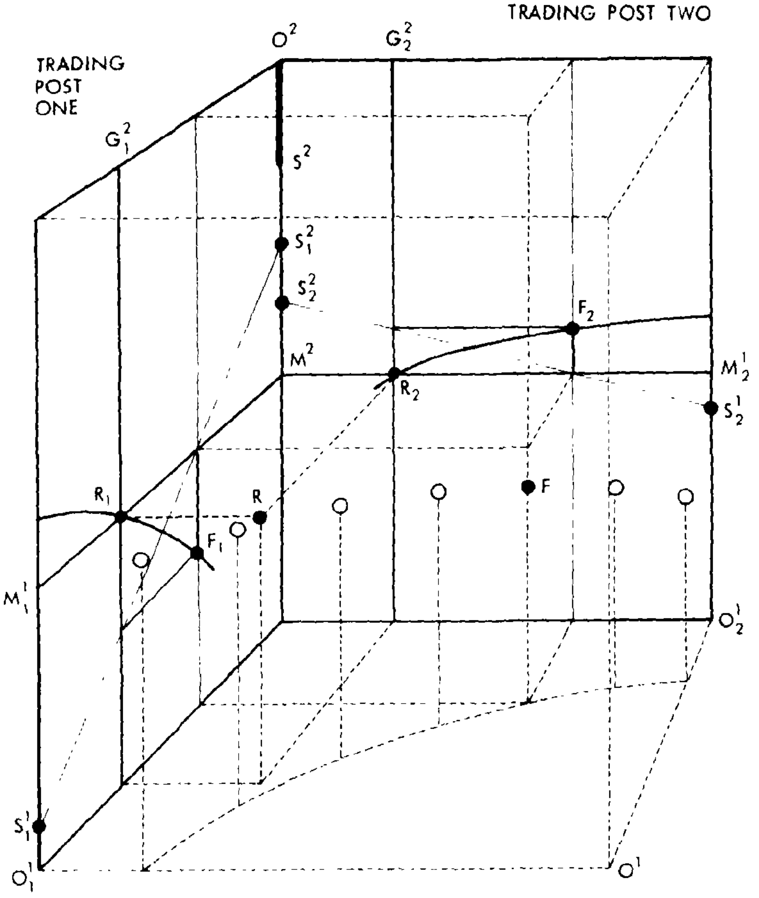


FIG 7. - Two trading posts

substitution effects, etc., among the different goods, so that the utility maps we superimpose on each Edgeworth box will depend on what is happening in the others. The second effect (as in the classical model) can only be visualized with hyperdimensional eyesight, but the first is still within reach of our diagrams.

Figure 7 illustrates a three-commodity situation. The two "boxes" are erected at right angles to each other, in the back of the three-dimensional figure (solid lines). The initial point, *R*, projects to R_1 in the first box and R_2 in the second.[24] We may conveniently regard *R* as the vector sum of M^2R_1 and M^2R_2. Trader 2's pair of bids b_1^2 and b_2^2 are shown at S_1^2 and S_2^2 and their sum at S^2. Thus $|M^2S^2| = |M^2S_1^2| + |M^2S_2^2|$. The unspent balance is represented by the vertical bar O^2S^2.

---

[24] The labels are generally consistent with the previous diagrams, but with subscripts added to distinguish the two trading posts.

Trader 1's bids are shown at S_1^1 and S_2^1. As S_1^1 varies over the permitted range $O_1^1M_1^1$, the outcome in the first marketplace sweeps out the curve through R_1 and F_1 in the usual way. Similarly, the curve through R_2 and F_2 describes his possibilities in the second marketplace. But his joint choice of S_1^1, S_2^1 is restricted by the spending limit

$$|M_1^1S_1^1| + |M_2^1S_2^1| \leq |M_1^1O_1^1| = |M_2^1O_2^1|.$$

The case actually illustrated has him spending most of his cash on good 1 and the rest on good 2. (Had he held some back, we would have shown it by a vertical bar at O^1.) The resulting final allocation is found by taking the vector sum of M^2F_1 and M^2F_2; this yields the point F.

Now suppose that Trader 1 changes his allocation of cash between goods 1 and 2, causing points S_1^1 and S_2^1 to move in opposite directions. The point F will then trace out a curve in space, which we have tried to suggest by the open dots perched above the curve's projection on the base. This projection on the base shows the tradeoff between goods 1 and 2 when a fixed amount of cash is bid by Trader 1; note that it is concave to his origin O^1. If Trader 1 bids less than the maximum allowed (strict inequality above), a curved triangular surface is generated, extending out from F and the open dots toward the viewer and rising to an apex directly above O^1. This locus (not shown) indicates ways in which Trader 1 can obtain more of good 3 at the expense of goods 1 and/or 2; it is concave to the origin O^1, and its projection is the large, roughly triangular region in the base of the diagram. The first trader's best response to the other trader's given strategy (S_1^2, S_2^2) is determined by the relationship of this surface to the indifference surfaces of u^1, in the three-dimensional commodity space with origin at O^1.

V. Replication

Inquiries into the behavior of economic models with large numbers of participants often make sweeping assumptions of symmetry, in the hope of keeping the models mathematically tractable and easy to visualize, while capturing at least some of the characteristic effects of large numbers. A favorite technique, involving a high but not total degree of symmetry, is called "replication." Imagine a basic economic system juxtaposed to a large number of identical replicas of itself; then take away all barriers, and form a "common market." Equivalently, assume that all traders in the full model are drawn from a small number of *types*, with an equal number of individuals of each type. Traders of the same type have identical endowments and identical tastes, but they are not constrained to act alike. That is, they are not programmed robots or members of a bloc or cartel but independent decision makers.[25]

The number of members of a type—the "replication number"—provides the modeler with a simple size parameter that he can vary without calling for additional data. Of course we do not intend to treat replication as though it were some kind of actual economic event, like a homogeneous population increase or an accretion of similar countries to a common market. Replication should be considered only as a technical device of comparative statics. As such, however, it has repeatedly proved its worth in explorations of the size effect (see Edgeworth 1881; Shubik 1959, 1968; Debreu and Scarf 1963, Shapley and Shubik 1966, 1969*b*; Debreu 1975; Owen 1975; Shapley 1975; etc.).

A. A Simple Case

The effect of replication on our model is illustrated in figure 8. We have $m = 1$ and $n = 2k$, where k is the replication number. The point S^1 represents the *average* of the bids of the k traders of type 1; similarly S^2 for type 2.[26] The point F therefore indicates the average final bundles for the two types.

Let us focus on a typical member of type 1. His own bid might be, say, at $\tilde{S}^1$. To discover his personal final bundle, $\tilde{F}$, we must draw the line through $\tilde{S}^1$ that is *parallel* to S^1S^2; we then locate $\tilde{F}$ in the usual way. Of course $\tilde{F}$ necessarily falls on the straight line through R and F, since all transactions take place at the same price.

Now suppose our trader changes his bid from $\tilde{S}^1$ to M^1. That is, he decides to buy nothing. This depresses the average bid for type 1, moving it from S^1 to T^1, where $|S^1T^1|/|\tilde{S}^1M^1| = 1/k$. The new price is the slope of T^1S^2, so our trader's new final bundle is $\tilde{A}^1$, determined by the line through M^1 parallel to T^1S^2. Similarly, at the other extreme, if he decides to bid the limit, the result is $\tilde{B}^1$, determined by the line through O^1 parallel to U^1S^2, where $|U^1S^1|/|\tilde{S}^1O^1| = 1/k$. We should not be surprised to see $\tilde{B}^1$ fall outside the box. This means merely that our trader happens to have enough "cash" to buy more than the combined initial bundles of one trader of each type.

The locus of possible final bundles for our variable trader is the curve $\tilde{A}^1R\tilde{F}\tilde{B}^1$, shown in figure 8 for the case $k = 4$. The flattening of the curve, as compared with the analogous curve A^1RFB^1 in figure 3 (which is the case $k = 1$), clearly shows the diminished influence of a single individual on price after replication. Indeed the price variation in the present case

---

[25] This distinction can be very important in the game-theory approach but is less important in a behavioristic theory. As we shall see, replication changes the NE but not the CE.

[26] More generally, if there were many types, then the analogue of S^2 would be not the average but $1/k$ times the sum of all bids from traders not of type 1.

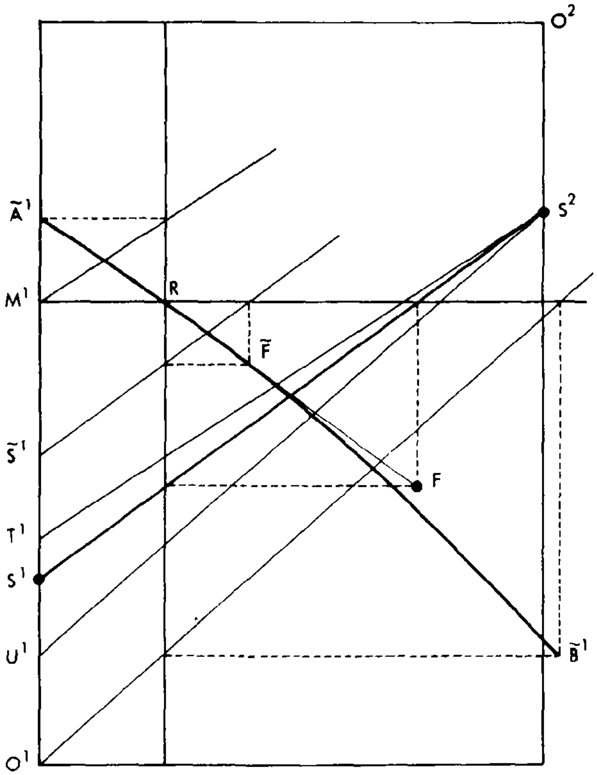


FIG. 8.—Price formation in a replicated market

is encompassed by the thin pencil of sloping lines within the angle $U^1S^2T^1$. In the limit, as $k \to \infty$, the curve becomes the fixed-price "budget set" of the classical CE model—but with one important difference: it is truncated at the $\tilde{B}^1$ endpoint instead of extending to the horizontal axis through O^1, in recognition of the limited quantity of the payment commodity available.[27]

B. The General Case

For the replicated form of the general model presented in Section II, we consider a market with T types of traders and k of each type; thus $n = kT$. The initial holding of good j by trader s of type t will be denoted

---

[27] It may be observed that the "conservative" line of credit for an individual shrinks to zero as $k \to \infty$. A simple calculation shows in fact that, in the present case, the amount of such credit is equal to $a_2^1a_2^2/[k(a_2^1 + a_2^2) - a_2^1]$.

a_j^{ts}, which we can shorten to a_j^t since it does not depend on s. The total amount of good j (summed over *both* t and s) will be denoted $\bar{a}_j$, as before. A typical strategy will be denoted $b^{ts} = (b_1^{ts}, \ldots, b_m^{ts})$. The defining equations of Section II then become

$$\begin{aligned} p_j &= \bar{b}_j/\bar{a}_j, && j = 1, \ldots, m; \\ x_j^{ts} &= b_j^{ts}/p_j, && j = 1, \ldots, m \text{ and } p_j \neq 0; \\ x_j^{ts} &= 0 && j = 1, \ldots, m \text{ and } p_j = 0; \end{aligned}$$

$$x_{m+1}^{ts} = a_{m+1}^{ts} - \sum_{j=1}^{m} b_j^{ts} + \sum_{j=1}^{m} a_j^{ts} p_j;$$

and

$$\Pi^{ts}(b^{11}, \ldots, b^{1k}, b^{21}, \ldots, \ldots, b^{nk}) = u^t(x_1^{ts}, \ldots, x_{m+1}^{ts}),$$

where $u^t(x) \equiv u^{ts}(x)$ is the common utility function for members of type t.

Suppose that a NE has been found in which all traders of the same type make the same bids:[28] $\hat{b}^{ts} = \hat{b}^t$, $s = 1, \ldots, k$. It is instructive to look at the equilibrium conditions for the "interior" case, that is, the case where all $\hat{b}_j^t$ are positive and $\sum_{j=1}^{m} \hat{b}_j^t < a_{m+1}^t$, $t = 1, \ldots, T$. Assuming that the concave functions $u^t(x)$ are differentiable, we let $u_j^t(x)$ denote their partial derivatives with respect to x_j. Then setting $du^t/db_j^{ts} = 0$ gives us

$$u_j^t(x^{ts}) \left\{\frac{1}{p_j} - \frac{b_j^{ts}}{(p_j)^2 \bar{a}_j}\right\} = u_{m+1}^t(x^{ts}) \left\{1 - \frac{a_j^{ts}}{\bar{a}_j}\right\},$$

which must hold at $b = \hat{b}$. Since prices are in terms of good $m + 1$ as numeraire, we can introduce $p_{m+1} \equiv 1$ and rewrite the condition for a symmetric, interior NE as follows:

$$\frac{u_j^t(\hat{x}^t)}{\hat{p}_j} \left\{1 - \frac{\hat{b}_j^t}{\bar{b}_j}\right\} = \frac{u_{m+1}^t(\hat{x}^t)}{\hat{p}_{m+1}} \left\{1 - \frac{a_j^t}{\bar{a}_j}\right\}, \qquad j = 1, \ldots, m,$$

where $\hat{p}$ and $\hat{x}$ are the prices and allocation corresponding to $\hat{b}$.

The expressions in curly brackets reveal the effect of oligopoly, but note that they lie between 1 and $1 - 1/k$. In the limit, therefore, these conditions for NE reduce to the conditions for CE, with the prices of the goods proportional to the marginal utilities of every trader. Thus we may state a convergence theorem, as follows:[29]

> *Theorem 2.* Assume that for infinitely many values of k the market has a symmetric, interior NE, and let $\hat{p}^{(k)}$ be the corresponding m-vector of prices. Let $\bar{p}$ be any limit point of the $\hat{p}^{(k)}$,

---

[28] The conditions given in Theorem 1 in Sec. IIIB are sufficient to guarantee the existence of such a symmetric NE. It is quite possible also to have nonsymmetric NEs.

[29] A somewhat more general form of this theorem will be found in Shapley (1976).

and define $\tilde{p}_{m+1} = 1$. Then the $m + 1$ prices $\tilde{p}_1, \ldots, \tilde{p}_m, \tilde{p}_{m+1}$ will be competitive for the market (for any value of k); that is, an allocation x will exist for which $\bar{\tilde{x}} = \bar{a}$ and, for each s and t, $\tilde{x}^{ts}$ maximizes $u^t(x^{ts})$ subject to $x^{ts} \geq 0$ and

$$\sum_{j=1}^{m+1} \tilde{p}_j(x_j^{ts} - a_j^{ts}) = 0.$$

It should be noted that the NE approaches the CE "from below," that is, through outcomes that are not in general Pareto optimal. This contrasts with the convergence of cooperative solutions like the core and the value, which are by definition Pareto optimal all the way (see, e.g., Debreu and Scarf 1963; Shapley and Shubik 1969*b*; Debreu 1975; Owen 1975; and Shapley 1975).

The type of convergence revealed in Theorem 2 depends crucially on having interior solutions, with no trader up against his spending limit. All must have enough cash (or credit, in an extended model), where the meaning of "enough" depends on the payment commodity's marginal utility relative to that of the other goods. Of course there can never be "enough" fiat money in this sense (in a one-period model), for the NE will always have everyone spending the limit and wishing he could spend more. If the CE is considered socially desirable (as one road to the Pareto optimum, if for no other reason), then a society whose trading system resembles our replicated model should make sure that the means of payment is something that is both widely available and generally desirable, either as a consumption item or as a means of payment in future periods.

VI. Variants and Extension

It is not difficult to modify the basic game so that the goods do not necessarily all pass through the market before consumption. A number of different considerations and problems arise, however, depending on just how this is done. We shall review some of them briefly here; for further details see Shapley (1976).

Suppose, first, that a trader can hold back any part of his initial endowment. A strategy for i is now a *pair* of m-dimensional vectors (b^i, q^i), with the b_j^i representing cash bids, as before, and the q_j^i representing quantities of goods sent to the respective trading posts. Thus the amounts $(a_j^i - q_j^i)$ are held back. Of course we require that

$$0 \leq q_j^i \leq a_j^i, \qquad j = 1, \ldots, m.$$

The prices are given by

$$p_j = \bar{b}_j / \bar{q}_j \qquad (\text{if } \bar{q}_j > 0),$$

and we add the convention that if $\bar{q}_j = 0$ any cash sent to the jth trading post is lost. Sales and purchases are calculated as before, but we must remember to add back the amounts $(a_j^i - q_j^i)$ before entering the utility function.

The nomogram in the Edgeworth box works essentially as before, except that the domain of strategies S^i is now the full rectangle $M^iRG^iO^i$ instead of the line segment M^iO^i (see fig. 2). It is easy to see that all outcomes in the northwest and southeast quadrants can now be reached without the aid of credit. But when there are more than two traders, we find the feasible set still sharply restricted by the requirement of trading at fixed prices, as discussed previously in Sections IIC and IVD.

Despite the naturalness of the "hold-back" option, it has one major disadvantage. It turns out that the enlarged strategy spaces cause the conditions that define the noncooperative equilibrium to be *under*determined, making the set of solutions usually infinite.[30] This circumstance leads us to search for ways to sharpen either the solution concept or the rules of the game, without entirely giving up the hold-back feature. The simplest expedient is to prohibit a trader from simultaneously buying *and* selling at the same trading post, by imposing the condition

$$b_j^i q_j^i = 0, \qquad j = 1, \dots, m.$$

The strategy domains at each trading post are now L-shaped, with a corner at the initial point R. Despite this corner, a trader in an active trading post should not experience any discontinuity in changing roles from buyer to seller, as his budget set is still the familiar, smooth hyperbola through R, as long as there are other traders on both sides of the market. The two-person case is rather trivial, however, there being no meaningful competition to motivate price formation.

This modification does restore some semblance of uniqueness to the NE, but it has its own drawback: the game is *not additive* over players. A coalition has options not available to an individual. Thus, if a trader could split himself into two legal persons, he could buy and sell simultaneously at the same trading post (e.g., in an attempt to stabilize the price). This suggests that the restriction $b_j^i q_j^i = 0$ might be unrealistic in some applications.

Any version of the rules that permits traders to stay out of a market has another peculiarity, namely, the possibility of trading posts that are completely inactive. Such a situation is very stable, as no trader would want to enter a market where $\bar{b}_j = \bar{q}_j = 0$. He would either lose his goods or money or, at best, have them returned to him to no advantage. Thus the "null" strategy for all traders is always an equilibrium point, in both of the models above. Moreover, we can arbitrarily declare any subset of trading posts to be inactive and solve the remaining subeconomy —any NE of that subeconomy will be also a NE of the economy as a whole.

---

[30] This multiplicity is not merely a matter of simultaneous buying and selling by the same trader that could simply be canceled out. Indeed, if at equilibrium Trader i is sending both goods and cash to the same trading post and if the price there is p_j, then he might consider decreasing both q_j^i and b_j^i, in the ratio of 1 to p_j. This would not change his final outcome (assuming that he does not spend the extra cash elsewhere) or the price p_j, but it would change the marginal cost of good j to the other traders and so destroy the equilibrium.

We should not completely rule out such inactive or partially inactive solutions as unrealistic on their face. It can plausibly be argued that the real world is full of "latent" markets awaiting discovery which can only become active by an act of faith on the part of the first entrants. But we are interested in active NE as well, and the possibility of inactive ones makes the main existence theorem, comparable to Theorem 1 above, considerably more delicate both to state and to prove. Sometimes markets are "legitimately" inactive, as when the traders start with a Pareto-optimal distribution. The resolution of the problem seems to lie in demanding that inactive trading posts have "virtual prices" at which all traders find it in their best interest neither to buy nor to sell (Shapley 1976).

A. Multiperiod Extensions

Many applications of our general approach can benefit from—and some will require—the adoption of a multiperiod framework. Specifically, we may mention the study of (*a*) credit and bankruptcy, (*b*) nonsymmetric information conditions, (*c*) uncertainty and insurance, (*d*) cyclical variation of endowments and the money market, (*e*) the derived utility of fiat money, (*f*) interest rates and inflation, and (*g*) the role of capital goods and ownership shares. Shubik and others have devised exploratory game models for most of these situations and have worked out a number of tutorial examples (see Shubik 1972, 1973, 1977; Shubik and Whitt 1973; and Dubey and Shubik 1976). There remains, however, much more that can and should be done.

There remain also some serious conceptual problems with the general mathematical theory for multistage models of this type. One kind of difficulty is already apparent if we try merely to extend our basic prototype, without added features, to two or more periods. We should first recall that a vital ingredient in the proof of Theorem 1 (the existence theorem) is the *concavity* of the payoff functions $\Pi^i(b^1, \ldots, b^n)$, with respect to their respective b^i. The natural way to attack the two-period case is to solve the second stage parametrically, as a function of the cash distribution at the end of the first stage. But then the traders' total payoffs will include a term that depends on this cash distribution, and there seems no way to assure that this term will be concave.

VII. Conclusion

In this paper, in an admittedly simplified and abstract setting, we have explored some elementary implications of an explicit market mechanism for the formation of price. We believe that this species of game-theoretic model, by being well defined independently of equilibrium conditions or behavioral assumptions (though capable of accommodating such conditions and assumptions) and by reflecting the decentralized decision making attainable through the use of a tangible money, is far more flexible as an investigatory and explanatory tool than the usual Walrasian model, with its ill-defined causal linkage between individual actions and the action of the system as a whole.

References

Arrow, K. J., and Debreu, G. "Existence of an Equilibrium for a Competitive Economy." *Econometrica* 22 (July 1954): 265–90.

Arrow, K. J., and Hahn, F. H. *General Competitive Analysis*. San Francisco: Holden Day, 1971.

Aumann, R. J., and Peleg, B. "Von Neumann-Morgenstern Solutions to Cooperative Games without Side Payments." *Bull. American Math. Soc.* 66 (May 1960): 173–79.

Aumann, R. J., and Shapley, L. S. *Values of Non-atomic Games*. Princeton, N.J.: Princeton Univ. Press, 1974.

Clower, R. "A Reconsideration of the Microfoundations of Monetary Theory." *Western Econ. J.* 5 (March 1967): 1–9.

Cournot, A. A. "Recherches sur les principes mathématiques de la théorie des richesses." Paris: Hachette, 1838. Translated by N. T. Bacon. New York: Macmillan, 1897.

Debreu, G. "A Social Equilibrium Existence Theorem." *Proc. Nat. Acad. Sci.* 38 (October 1952): 886–93.

———. "Economies with a Finite Set of Equilibria." *Econometrica* 38 (May 1970): 387–92.

———. "The Rate of Convergence of an Economy." *J. Math. Econ.* 2 (March 1975): 1–7.

Debreu, G., and Scarf, H. E. "A Limit Theorem on the Core of an Economy." *Internat. Econ. Rev.* 4 (September 1963): 253–46.

Dubey, P., and Shubik, M. "Trade and Prices in a Closed Economy with Exogenous Uncertainty, Different Levels of Information, Money and Special Futures Markets." Cowles Foundation Discussion Paper no. 410, Yale Univ., January 1976.

Edgeworth, F. Y. *Mathematical Psychics: An Essay on the Application of Mathematics to the Moral Sciences*. London: Kegan Paul, 1881.

Hildenbrand, W. *Core and Equilibria of a Large Economy*. Princeton, N.J.: Princeton Univ. Press, 1974.

Jones, R. A. "The Origin and Development of Media of Exchange." *J.P.E.* 84, no. 4, pt. 1 (August 1976): 757–75.

Levitan, R. E., and Shubik, M. "Noncooperative Equilibria and Strategy Spaces in an Oligopolistic Market." In *Differential Games and Related Topics*, edited by H. Kuhn and G. Szegö. Amsterdam: North-Holland, 1971.

———. "Price Duopoly and Capacity Constraints." *Internat. Econ. Rev.* 13 (February 1972): 111–22.

Mañas, M. "A Linear Oligopoly Game." *Econometrica* 40 (September 1972): 917–22.

Nash, J. F., Jr. "Noncooperative Games." *Ann. Math.* 54 (September 1951): 286–95.

Ostroy, J. M., and Starr, R. M. "Money and the Decentralization of Exchange." *Econometrica* 42 (November 1974): 1093–1113.

Owen, G. "On the Core of Linear Production Games." *Math. Programming* 9 (December 1975): 358–70.

Patinkin, D. *Money, Interest, and Prices: An Integration of Monetary and Value Theory.* Evanston, Ill.: Row, Peterson, 1956.

Postlewaite, A., and Schmeidler, D. "Approximate Efficiency of Non-Walrasian Nash Equilibria." Discussion Paper, Univ. Illinois at Urbana-Champaign, October 1975.

Samuelson, P. A. *Foundations of Economic Analysis.* Cambridge, Mass.: Harvard Univ. Press, 1947.

Scarf, H. E. "The Core of an *n* Person Game." *Econometrica* 35 (January 1967): 50–69.

———. *The Computation of Economic Equilibria.* New Haven, Conn.: Yale Univ. Press, 1973.

Shapley, L. S. "An Example of a Slow Converging Core." *Internat. Econ. Rev.* 16 (June 1975): 345–51.

———. "Noncooperative General Exchange." In *Theory and Measurement of Economic Externalities*, edited by S. A. Y. Lin. New York: Academic Press, 1976.

Shapley, L. S., and Shubik, M. "Quasi-Cores in a Monetary Economy with Nonconvex Preferences." *Econometrica* 34 (October 1966): 805–27.

———. "On Market Games." *J. Econ. Theory* 1 (June 1969): 9–25. (*a*)

———. "Pure Competition, Coalitional Power, and Fair Division." *Internat. Econ. Rev.* 10 (October 1969): 337–62. (*b*)

———. "The Assignment Game. I. The Core." *Internat. J. Game Theory* 1, no. 2 (1972): 111–30.

———. "Characteristic Function, Core, and Stable Sets." Rand Corporation Report R-904/6, Santa Monica, Calif., July 1973.

———. "Competitive Outcomes in the Cores of Market Games." *Internat. J. Game Theory* 4, no. 4 (1976): 229–37.

Shubik, M. "Edgeworth Market Games." In *Contributions to the Theory of Games*, edited by R. D. Luce and A. W. Tucker. Vol. 4. Annals of Mathematics Study no. 40. Princeton, N.J.: Princeton Univ. Press, 1959.

———. "Extended Edgeworth Bargaining Games and Competitive Equilibrium." *Metroeconomica* 20 (September–December 1968): 299–312.

———. "Pecuniary Externalities: A Game Theoretic Analysis." *A.E.R.* 61 (September 1971): 713–18.

———. "A Theory of Money and Financial Institutions: Fiat Money and Noncooperative Equilibrium in a Closed Economy." *Internat. J. Game Theory* 1, no. 4 (1972): 243–68.

———. "Commodity Money, Oligopoly, Credit and Bankruptcy in a General Equilibrium Model." *Western Econ. J.* 11 (March 1973): 24–38.

———. "A Dynamic Economy with Fiat Money, without Banking, but with Ownership Claims to Production Goods." In *Mélanges F. Perroux*, edited by H. Bartoli. Paris, 1977, in press.

Shubik, M., and Whitt, W. "Fiat Money in an Economy with One Nondurable Good and No Credit." In *Topics in Differential Games*, edited by A. Blaquière. Amsterdam: North-Holland, 1973.

Telser, L. G. *Competition, Collusion, and Game Theory.* Chicago: Aldine-Atherton, 1972.

Viner, J. "Cost Curves and Supply Curves." *Zeitschrift Nationalökonomie* 3 (1931): 23-46.

von Neumann, J. "Zur Theorie der Gesellschaftsspiele." *Math. Ann.* 100 (1928): 295-320. (Or see the English translation by Sonya Bargmann in *Contributions to the Theory of Games*, edited by R. D. Luce and A. W. Tucker. Vol. 4. Annals of Mathematics Study no. 40. Princeton, N.J.: Princeton Univ. Press, 1959.)

Organized Futures Markets: Costs and Benefits

Lester G. Telser and
Harlow N. Higinbotham
University of Chicago

A futures contract is to a forward contract as payment in currency is to payment by check. An organized market facilitates trade among strangers. Such a market trades a standardized contract under appropriate rules. The equilibrium distribution of market clearing prices is asymptotically normal with a standard deviation that varies inversely with the volume of trade, given underlying supply and demand conditions. Empirical relations giving the commission and margin per contract as a function of the volume of trade and outstanding commitments for 23 commodities support the theory. Also, comparisons of pertinent aspects of 51 commodities divided into active, less active, and dormant groups are consistent with the theory.

I. Introduction

Of all of the hundreds of thousands of commodities in the economy only a few have ever been traded on an organized futures market. This is a puzzle. The basic idea we use to give an answer to this puzzle draws on the theory of money. An organized futures market creates a medium of exchange, a futures contract, with many of the attributes of money. A futures contract facilitates trade in the commodity in the same way that the use of money in exchange has advantages over trade by barter under normal conditions. Nor is this all. A futures contract is a temporary abode

---

We gratefully acknowledge the financial support of the National Science Foundation. Members of the Industrial Organization Workshop and of the Seminar in Applied Price Theory of the University of Chicago gave helpful comments on earlier drafts. Charles Cox, Edward Lazear, George Stigler, and an anonymous referee deserve our thanks. All remaining errors and defects are our sole responsibility.

[*Journal of Political Economy*, 1977, vol. 85, no. 5]
© 1977 by The University of Chicago. All rights reserved.

of purchasing power in terms of the commodity. It is these aspects of a futures contract and not the more common view that futures contracts enable hedgers to avoid risks that explain the benefits of an organized futures market.

A transaction in a physical commodity in the real world has many unique characteristics stemming partly from the identities of the parties to the transaction, their reliability, credit worthiness, promptness, honesty, and flexibility; the qualities of the good; and the circumstances of the trade. These particulars make a transaction in the physical commodity resemble a barter trade. The two parties in a mutually acceptable trade have costs like those that arise in the double coincidence of trade by barter. The introduction of a standard futures contract by an organized futures market creates a financial instrument that can be traded without knowing the actual identity of the two parties in the transaction. The seller of the futures contract incurs a liability to the organized futures market, and the buyer acquires an asset from this market. Neither need have concern about the integrity of the other in the same sense that one who accepts a $10 banknote in payment for something need not worry about the credit rating of the buyer. This argument implies that any commodity not made to order can benefit from the introduction of an organized futures market. The latter has a cost. Therefore, the use of this commercial invention, the futures contract, appears only for those commodities where the benefits outweigh the costs.

The development of a formal theory of these costs and benefits begins with a stochastic model of market clearing prices which has its roots in the economics of information and search. This model shows that the distribution of market clearing prices is asymptotically normal. The standard deviation of the limiting normal distribution depends on certain properties of the underlying schedules of bids and offers. Also, the realized market clearing price approaches the asymptotic market clearing price in probability as the number of traders increases. This theory is a marriage of economics and probability.

The standard deviation of the distribution of market clearing prices measures the liquidity of the market. There is a cost of lowering the standard deviation and there are benefits. The optimal amount of liquidity takes both the costs and benefits into account and results in some positive value of the standard deviation.

Although the standard deviation is an important element in our theory and seems to be a roundabout way of introducing risk, our emphasis is different and leads to different implications than a theory of futures markets based on the concept of reducing risk. An organized futures market is not necessary in order to obtain the advantages of hedging. These can result from a system of long-term contracts. However, ar organized futures market can reduce the costs of moral hazards by th introduction of a standard contract. A long-term contract between two parties depends on their integrity. In a situation with many participants, a standard contract with the backing of the organized market has less moral hazard. Ours is not a theory of why there may be advantages in long-term contracts; it is a theory of the net benefit of an organized exchange. There is some common ground since a long-term contract and a futures contract both refer to dates in the future. Hence an organized futures market is more likely to occur in commodities where the timing of transactions is more important. This only gives a set of necessary conditions, and it does not furnish sufficient conditions for the emergence of an organized futures market.

More generally, this theory asserts that a necessary feature of an organized market is a standard contract traded in that market. This standard contract need not be a futures contract. It may be a certificate that allows the bearer to obtain on demand a specified quantity of a good. Certain forms of certified warehouse receipts, gold and silver certificates, and similar financial instruments, such as shares of stock, have these properties. If the standard contract represents a physical good, then some legal entity is liable for fulfilling the terms of the contract on demand. This person would hold sufficient reserves to maintain the probability of default at an optimal level. Reserves equal to a fraction of the outstanding commitments of the standard contracts may suffice. These standard contracts may circulate among the traders at market determined prices, and the participants may never wish to convert their standard contracts into the physical good. This is plain in the case of shares of stock since rarely do the shareholders wish to liquidate the corporation.

This argument implies that the number and identity of the participants in a potential organized market is important. If this number is small, if the potential participants know each other, and if there is little turnover of these participants, then they may not need a standard contract to deal with each other. If there is a large number of potential participants and a rapid turnover of these traders, then this raises the benefit of a standard contract as well as the benefit of having an organized exchange. The members of the exchange meet certain requirements so that those who deal with them can rely on their integrity. The members of the organized exchange also face the problem that their clients may lack integrity. That is, as the number of potential participants in an organized market increases, the cost of having them all trade on their own account in the organized market may rise more than proportionally. This is not only because of the increasing congestion this would cause but also because there may be increasing costs of having one organization certify all of the potential traders.

Even without an organized market in a standard contract the traders may have confidence in each other if each is a legal representative of well-known principals. This is an important consideration when the trade is in financial instruments that represent actual goods or assets instead of trade in the physical goods or assets themselves. The latter does occur in, say, a flea market. Even if the buyer is a good judge of the commodity before him in the flea market, there is still the problem that he must know that the seller is the legal owner of the good or is the agent of the legal owner. These are important problems in the real world that are often overlooked in some formal theories of market exchange.

The theory herein owes a debt to some important contributions by others. Most notable is Holbrook Working, who always put the emphasis on the functions of a futures market aside from its advantages to hedgers. In some of his work he came close to regarding a futures contract as a temporary abode of purchasing power in the physical commodity (1953, 1967). H. S. Houthakker also points out the similarities between money and futures contracts in his 1959 article. In contrast, most of the economists who have written on the subject, beginning with Marshall's *Industry and Trade* (1920), follow the convention of focusing their attention on hedging and speculation, which loses sight of the more fundamental properties of an organized exchange.

There is another important aspect of the subject that is peripheral to our main interest, and we do not discuss it. This refers to the actual rules of the exchange and the operation of the market. This topic deserves the close attention of economists because it deals with the practical problem of creating a set of conditions that can make a real market function according to the theoretical model of a perfectly competitive market. Informed observers of actual futures markets know that this is a difficult task. An organized futures market cannot survive unless it does approximate a perfectly competitive market.

II. Properties of an Organized Exchange

Before presenting a theory of the costs and benefits of an organized futures market, it is helpful to describe some of its important properties. Trade occurs in one physical place, the floor of the exchange, durin specified hours, called the trading session. The traders cry out bids an offers, making a bilateral auction market. Only members of the exchang may trade on the floor, and no member of the exchange may mak transactions off the floor of the exchange. Members of the exchang either trade for themselves or under the instructions of their clients wh are typically not members of the exchange. Many organized exchang are responsible for ensuring that the terms of a transaction are fulfill The two parties in a transaction agree on price and quantity of the giv futures contract. All other terms of the contract are specified by the rul of the exchange.

The exchange determines the number of members who may trade on the floor. A membership is often called a seat on the exchange. A member may sell his seat to a nonmember. However, this is subject to some control by the exchange. The exchange investigates the character of potential members, and it may refuse permission to a potential buyer of a seat. The price of the seat is mutually agreed upon by the seller and buyer and is not subject to control by the exchange.

The exchange can discipline its members by imposing fines, suspending their trading privileges, or by expelling them. The oldest organized futures exchanges, such as the Chicago Board of Trade, have evolved elaborate rules as a result of their long experience. These rules intend to give those who trade on the exchange confidence in the reliability of the transaction executed on the exchange. Members of the exchange must execute the orders of the public, that is, nonmembers, before they execute their own trades. Exchange members who make fictitious trades are subject to penalties. Such fictitious trades intend to record prices that will mislead others. The exchange defines the terms of a contract such that all contracts of a given class are perfect substitutes and such that the validity of a transaction in that contract does not depend on the identity of the principals. Thus, a standard contract for the delivery of 5,000 bushels of wheat in July 1976 executed on the floor of the Chicago Board of Trade in January 1976-at a price mutually agreeable to the parties is as well defined as currency. A futures contract is to a forward contract as currency is to a check drawn against a demand deposit in a commercial bank. The validity of a genuine $10 bill, one printed in the U.S. Bureau of Engraving, does not depend on who offers it in payment, while the validity of a $10 check depends on the identity of the person who writes or presents it and on the identity of the bank. Similarly, the validity of a July futures contract on the Chicago Board of Trade is as good as the faith and credit of the Chicago Board of Trade, while the validity of a forward contract for July delivery depends on the integrity of the buyer and seller in the transaction.

A forward contract shares with a futures contract the important property that the buyer and the seller agree in the present on the terms of a transaction that will be completed at a specified time in the future. The important distinction between the two kinds of contracts lies in this. In a forward contract the actual identity of the buyer and seller is important. Neither has recourse in case of dispute to a third party other than a court of law. The validity of the forward contract depends on the good faith of the two parties themselves. A futures contract has a third party, the organized exchange or its designated representative, that guarantees the validity of the contract and will enforce the terms. When A sells forward to B, the consummation of the transaction depends on their honesty. When A sells a futures contract to B, A incurs a liability to the organized exchange and B acquires an asset from the organized exchange. The

As the number of traders in the market increases indefinitely, the market clearing price approaches the solution given by (3).

Under mild assumptions about the cumulative densities, $U(\cdot)$ and $V(\cdot)$, this theory implies that the distribution of the market clearing prices is asymptotically normal. This result also holds even if individual traders each make bids or offers for more than one unit of the good, provided each trader has a negatively sloping excess demand and is small, in an appropriate sense, relative to the whole market. Since the details are available elsewhere (Telser, forthcoming), the proof here is brief. The population quantile of offers that can expect acceptance at the asymptotic market clearing price p_e is $U(p_e)$. Given a random sample of m offers, calculate the sample fractile, which is some price, corresponding to the population fractile $U(p_e)$. If (3) has a unique solution and if $U(p)$ has a derivative in a neighborhood of p_e and at p_e, then the random variable $\sqrt{m}(p_m - p_e)$ converges in distribution to normality with mean zero and standard deviation $\sigma(p_m)$, where

$$\sigma(p_m) = \{1/u(p_e)\}\{\sqrt{[U(p_e)(1 - U(p_e))/m]}\}. \tag{4}$$

Similarly, for a random sample of n bids, the random variable $\sqrt{n}(p_n - p_e)$ converges in distribution to normality with mean zero and standard deviation $\sigma(p_n)$ where the expression for $\sigma(p_n)$ is the same as for $\sigma(p_m)$, after making the appropriate substitutions, $1 - V$ for U, V for $1 - U$, v for u, and n for m. By hypothesis, the samples of bids and offers are independent. Therefore, the distribution of market clearing prices is also asymptotically normal with mean zero and standard deviation $\sigma(p_{m,n})$, where

$$\sigma^2(p_{m,n}) = \sigma^2(p_m) + \sigma^2(p_n). \tag{5}$$

Several important propositions follow from this analysis. First, the standard deviation of the distribution of market clearing prices is a decreasing function of the square root of the number of transactions. Second, the standard deviations are lower the larger the slopes of the cumulative densities. The slopes measure the homogeneity of the offers or bids in the population. If all offers were the same in the population then $dU(p_e)/dp = \infty$, which corresponds to an infinitely elastic supply. Consequently, (4) asserts, $\sigma(p_m) = 0$. If all offers in the population were different so that dU/dp approaches zero, then $U(p)$ would approach an improper uniform distribution on the halfline $[0, \infty]$ and $\sigma(p_m)$ would increase indefinitely. Therefore, the homogeneity of offers varies inversely with dU/dp. Third, the standard deviation has a maximum if $V(p_e) = U(p_e) = \frac{1}{2}$.[1]

---

[1] The effect of heterogeneity on the standard deviation is given as follows. Assume there are n normal distributions, denoted by N_i, $i = 1, \ldots, n$. Let α_i be the probability of a random drawing from N_i. Let μ_i denote the mean of N_i and σ_i the standard deviation of

According to this theory the standard deviation of the distribution of market clearing prices approaches zero as the number of potential transactions per unit time increases. A given batch of transactions has one market clearing price at which mutually acceptable trades can occur. It is the size of this batch that determines $\sigma(p_{m,n})$. The cost is an increasing function of m and n and there is also a benefit. The optimal volume of trade depends on both. The analysis of costs and benefits in the next section determines the optimal value of σ and shows that it is positive. In this way we obtain a theory of the equilibrium price distribution with both the mean and σ explained by the theory.

IV. A Theory of the Net Benefit of an Organized Futures Market

The preceding section shows that the distribution of market clearing prices is asymptotically normal with a standard deviation, σ, given by

$$\sigma = \alpha K x^{-1/2}, \tag{6}$$

where x is the volume of trade corresponding to m and n in the preceding section, K a parameter that depends on the underlying supply and demand for the commodity, and α represents certain exogenous factors that we consider in more detail below. For present purposes it is convenient to write (6) in a more general form:

$$\sigma = g(x, \alpha), \tag{7}$$

N_i. Let X denote a random variable. Then

$$\mu = EX = \sum_i E(X:N_i) \Pr(N_i) = \sum_i \alpha_i \mu_i. \tag{i}$$

Also,

$$EX^2 = \sum_i E(X^2:N_i) \Pr(N_i) = \sum_i \alpha_i(\sigma_i^2 + \mu_i^2), \tag{ii}$$

$E(X^2:N_i) = \sigma_i^2 + \mu_i^2$. Since $\text{var } X = EX^2 - (EX)^2$, it follows that

$$\text{var } X = \sum \alpha_i(\sigma_i^2 + \mu_i^2) - (\sum \alpha_i \mu_i)^2. \tag{iii}$$

However, $\sum \alpha_i(\mu_i - \mu)^2 = \sum \alpha_i \mu_i^2 - \mu^2$, so that (iii) becomes

$$\text{var } X = \sum \alpha_i \sigma_i^2 + \sum \alpha_i(\mu_i - \mu)^2. \tag{iv}$$

If $\sigma_i^2 = \sigma^2$ for all i then (iv) reduces to

$$\text{var } X = \sigma^2 + \sum \alpha_i(\mu_i - \mu)^2. \tag{v}$$

Therefore,

$$\text{var } X - \text{var } X_i = \sum \alpha_i(\mu_i - \mu)^2 \geq 0, \tag{vi}$$

with equality if and only if $\mu_i = \mu$ for all i. This shows that heterogeneity raises the standard deviation of the distribution of the random variable X. For the application to the distribution of market clearing prices,

$$\sigma_i^2 = F_i(1 - F_i)/[m_i(\partial F_i/\partial p)^2], \tag{vii}$$

where m_i denotes the size of the random sample from N_i. Let p_e^i denote the asymptotic market clearing price in market i. If $F_i(p_e^i) \neq F_j(p_e^j)$ $i \neq j$, then there are differences both with respect to μ_i and with respect to σ_i.

and, in conformity with (6), $g_x < 0$, $g_\alpha > 0$, and $g_{\alpha x} < 0$. We assume that the same function $g(\cdot)$ applies to all commodities and that differences among commodities express themselves via different values of α.

The total cost function of having and operating an organized market is

$$c = h(x, X, \sigma), \tag{8}$$

where X is the size of the outstanding commitment of futures contracts, called the open interest. The presence of x, the volume of trade, as an argument of this total cost function needs no discussion. Assume $h_x > 0$. The open commitment, X, enters for two reasons. First, there is a cost of record keeping. A buyer and seller in the market agree on the price and quantity. This transaction creates a financial instrument on the balance sheet of the futures market, or, more exactly, its clearinghouse. The seller of a futures contract has a liability to the clearinghouse and the buyer has an asset of the clearinghouse. The clearinghouse must keep track of the identity of its debtors, called the shorts, and its creditors, the longs. This imposes a cost that is an increasing function of X. Some of the trade on an organized market comes from outsiders who are not members of the exchange. These trades go through the exchange members who carry out the orders of their customers. Hence the exchange members incur costs as a result of the actions they take on behalf of their customers. These costs depend on x, X, and σ and are included in the function $h(\cdot)$. Accordingly, $h_X > 0$.

The total cost also depends on both X and σ because an exchange member is liable to the clearinghouse for the commitments of his customers. The commodity price may change adversely to the customer of the exchange member. This can impose a loss on the exchange member although he can liquidate the customer's commitment in order to limit the loss. The exchange member normally requires a security deposit, called the margin, from his customer in order to protect himself against loss. This gives a cost per futures contract to the customer that is proportional to the margin per contract. If the only traders on the floor of the exchange were principals then it would not be necessary to have margins and this component of the total cost could be avoided. The higher is σ, the larger is the margin per unit of X and the higher is the total cost. Therefore, $h_X > 0$, $h_\sigma > 0$, and $h_{X\sigma} > 0$.

The net marginal cost with respect to volume is dh/dx given as follows:

$$dh/dx = h_x + h_\sigma g_x. \tag{9}$$

Since the theory stipulates $h_x > 0$, $h_\sigma > 0$, and $g_x < 0$, the net marginal cost can be negative. If the net marginal cost is positive, then, since $g_x < 0$, dh/dx can be a U-shaped function of x. It is more convenient to return to the analysis of this function in conjunction with the discussion of the equilibrium conditions given below.

The total benefit, denoted by b, is an increasing function of x and X. Since σ is the standard deviation of the distribution of market clearing prices, if it enters the benefit function at all, it should do so as a "bad." That is, an increase in σ lowers the total benefit. There would be little point in making σ an argument in the benefit function because this would simply complicate the algebra without yielding new results. The theory does assume that the total cost is an increasing function of σ. Consequently, as it stands, the net benefit, which is the total benefit minus the total cost, is a decreasing function of σ. This means that σ now enters the net benefit function as a bad.

There is another separate issue related to price variability. One reason often given for having an organized futures market is to enable hedging by dealers in the actual commodity. Someone who has a commitment in the actual good, according to this argument, can take an offsetting position in commodity futures contracts and thereby reduce the effects of a change in the price of the commodity. The price risk per unit of the commodity is an increasing function of the unpredictable price variability, denoted by β. Therefore, the benefit of an organized futures market is an increasing function of β. Write the benefit function as follows:

$$b = f(x, X, N, \beta). \tag{10}$$

Assume that f_x, f_X, and f_β are all positive. Now if β represents the unpredictable price variability, it should also affect σ, which is the standard deviation of the distribution of market clearing prices. Hence we modify (7) and write

$$\sigma = g(x, \alpha, \beta), \qquad g_\alpha > 0 \text{ and } g_\beta > 0. \tag{7'}$$

This formulation distinguishes between those factors as given by the parameter α that affect σ and those which affect price variability over longer periods of time as represented by β. For instance, seasonal variation can raise σ and does not necessarily affect β (see n. 1). This formulation shows that the effect of a rise of β on the net benefit is indeterminate because both the cost and the benefit are increasing functions of β.

The parameter N in the benefit function represents the number of traders who can benefit from the organized market. Recall that we regard a futures contract as analogous to a form of money that facilitates trade in the commodity. A futures contract corresponds to trade with a common medium of exchange. Since a common medium of exchange is more useful the larger the number of participants, the total benefit increases with $N, f_N > 0$. Also, let $f_{xN} > 0$ and $f_{XN} > 0$.

Assume that the equilibrium values of x and X give the maximal net benefit

$$J(x, X) = b - c, \tag{11}$$

and assume that the observed values of x and X are the equilibrium values. A sufficient condition for the existence of an equilibrium is that the net benefit function, $J(x, X)$, is strongly concave in x and X. Hence the equilibrium is given by the necessary conditions for a maximum of the net benefit subject to the nonnegativity constraints, x and $X \geq 0$. We obtain

$$df/dx - dh/dx \leq 0 \quad \text{and} \quad df/dX - dh/dX \leq 0. \tag{12}$$

The terms df/dx and df/dX giving the marginal benefit with respect to the volume of trade and open interest correspond to the demand for the services of an organized exchange. The terms dh/dx and dh/dX giving the marginal cost with respect to the volume of trade and open commitments correspond to the supply of services of an organized exchange. If the equilibrium values of x and X are positive, there is equality in both equilibrium conditions of (12). The two conditions are independent although one refers to x and the other to X, and under suitable conditions x is the rate of change of X. One can have $X = 0$ and $x > 0$, giving the implication of an upward jump of X. One may have $X > 0$ and $x = 0$ so that although there are positive outstanding commitments, there happens to be no trade. Finally, if the equilibrium values of x and X are both zero there is no organized market in the good. This theory states the absence of an organized market as a corner solution of the equilibrium conditions.

There is the complication that the net marginal cost, dh/dx, is not necessarily monotonic. Under the hypothesis that the net benefit function is strongly concave and that the benefit function $f(x, X, N, \beta)$ is concave in the first two arguments, there are definite restrictions on the properties of dh/dx. Thus, dh/dx must be increasing wherever it is negative. If dh/dx is positive and decreasing, then it must decrease less rapidly than df/dx at corresponding values of x.

Let t denote the commission per transaction. The theory asserts that in equilibrium with $x > 0$ the commission per transaction satisfies the equation

$$t = dh/dx = df/dx. \tag{13}$$

In practice the situation is more complicated. First, the members of the exchange determine the smallest unit that may be traded and the minimum allowable price change per contract. For example, a wheat futures contract is 5,000 bushels and the minimum allowable price change per contract is \$12.50 on the Chicago Board of Trade. There are floor traders on the exchange who do not carry an open position overnight and who make many trades during a trading session. Their activities may increase x and lower σ. The return to these traders depends on their cost and is a function of the minimum quantities that may be traded and the minimum allowable price change. The marginal cost also depends on the rate of arrival of information, the level of trading activity, and the distribution of the sizes of the transactions.

There is another important aspect of an organized market. There are traders and brokers who specialize in the commodity and who furnish an inventory of their services during a trading session almost independently of the volume of trade during the session. The number of these specialists in the long run varies directly with the expected volume of trade. For an inactive commodity the commission per unit transaction may include a fixed component that remunerates the specialists for the services they provide by their continued presence on the floor even when trade is light. As the expected volume of trade rises relative to the stock of outstanding contracts, this phenomenon diminishes in importance.

Let m denote the margin per contract. Assume that the cost of the margin is proportional to m. In equilibrium with $X > 0$, the margin per contract, up to the factor of proportionality to represent the cost, satisfies the equation

$$m = df/dX = dh/dX. \tag{14}$$

In practice the size of the margin depends on the nature of the commitment, whether it is speculative, a straddle, a spread, or a hedge. It also depends on the exchange rule that determines the maximum allowable change in the price per contract during a single trading session. Trade during a session cannot continue unless price changes remain within prescribed limits. As a result there is a relation between the margin per contract and the maximum allowable price change per contract. Recall that the members of the exchange often act on the instructions of their customers and that many of the principals do not directly participate in the trade on the floor of the exchange. This relation between the margin per contract and the maximum allowable price change per contract gives all of the brokers time to consult their customers to determine whether the customers wish to continue their commitments in the face of adverse price change by furnishing additional margin. Otherwise, a broker would close out the commitment of his customer. The larger is the maximal allowable price change during a trading session the larger is the margin that the broker demands from his customer. The margin is between one and two times the daily maximal allowable price change per contract.[2]

---

[2] For the 17 commodities in the sample of actively traded goods, the simple correlation between the maximal allowable daily price change per contract and the margin per contract is 0.812. The margin increases by $1.972 per contract for each $1 increase in the maximal allowable daily price change per contract. For the 23-commodity sample, the simple correlation is 0.796 and the slope is different. For each $1.00 increase in the maximal daily allowable price change per contract, the margin rises by $1.744. The 23-commodity sample includes six less actively traded ones. These results show that the increment in margin per dollar increment in daily price limit is lower for the less actively traded commodities.

This argument implies that in those organized markets where the principals trade directly it would be unnecessary to stop trading so that there would be no maximum allowable price change. In such markets margins would also not be needed.

We study the properties of this theory in two different ways. First, we see how changes in the values of the exogenous parameters, α, β, and N affect the equilibrium values of x and X. Second, we study how the margin and the commission depend on these exogenous parameters. We maintain the hypothesis that the functional forms of the benefit and cost functions are the same for all commodities and that differences among the commodities express themselves via the values of the exogenous parameters.

Assume positive values of the initial equilibrium of x and X so that the equilibrium conditions given by (12) are equalities. Write the first-order conditions of the maximum of $J = b - c$ in the form $J_x = J_X = 0$. Use the symbol z to denote one of the exogenous variables, α, β, or N. Differentiate the equilibrium conditions with respect to z and obtain the equations as follows:

$$\begin{bmatrix} J_{xx} & J_{xX} \\ J_{Xx} & J_{XX} \end{bmatrix} \begin{bmatrix} dx/dz \\ dX/dz \end{bmatrix} + \begin{bmatrix} J_{zx} \\ J_{zX} \end{bmatrix} = 0. \tag{15}$$

Call the matrix on the left-hand side of (15) M. The hypothesis that J is strongly concave implies that M is a negative definite matrix. Hence multiplication of the expression in (15) on the left by the row vector $[dx/dz \;\; dX/dz]$ gives

$$(dx/dz)J_{zx} + (dX/dz)J_{zX} > 0. \tag{16}$$

If $z = \beta$, then $J_{zx} = J_{\beta x}$ and $J_{zX} = J_{\beta X}$. Consequently,

$$\begin{cases} J_{\beta x} = f_{x\beta} - h_{x\sigma}g_{\beta} - h_{\sigma}g_{\beta x} - h_{\sigma\sigma}g_{\beta}g_{x}, \\ J_{\beta X} = f_{X\beta} - h_{X\sigma}g_{\beta}. \end{cases} \tag{17}$$

Both of these expressions are of an indeterminate sign. Therefore, (16) does not fix the signs of $dx/d\beta$ and $dX/d\beta$. This is an important result because it means that a rise in exogenous price variability represented by the parameter β has an effect on x and X that depends on the actual numerical values of the terms in (15) and that knowledge of the signs of the components is not sufficient to predict the effects. The same conclusion applies to α since

$$J_{\alpha x} = -h_{x\sigma}g_{\alpha} - h_{\sigma\sigma}g_{\alpha}g_{x}, \tag{18}$$

$$J_{\alpha X} = -h_{X\sigma}g_{\alpha} < 0. \tag{19}$$

The indeterminacy of the sign of $J_{\alpha x}$ is due to the negative value of g_x. As we shall see from the empirical results, the figures give strong support to the view that $g_x < 0$.

Take $z = N$. Then

$$J_{Nx} = f_{xN} > 0 \qquad \text{and} \qquad J_{NX} = f_{XN} > 0. \tag{20}$$

It is now convenient to write out the solution of (15) for $z = N$. We obtain

$$dx/dN = (-J_{XX} f_{xN} + J_{Xx} f_{XN})/\det M, \tag{21}$$

$$dX/dN = (J_{Xx} f_{xN} - J_{xx} f_{XN})/\det M. \tag{22}$$

The only terms with an open sign in (21) and (22) involve J_{Xx}. If J_{Xx} is positive, then both dx/dN and dX/dN must be positive. If so, the larger the number of potential participants in the market the larger is the equilibrium volume of trade and equilibrium open interest:

$$J_{Xx} = f_{Xx} - h_{Xx} - h_{\sigma X} g_x. \tag{23}$$

Assume X and x are complements in the benefit function so that $f_{Xx} > 0$. By the preceding argument, $h_{\sigma X} > 0$ and $g_x < 0$. Hence two of the three terms in J_{Xx} would be positive. Even if $J_{Xx} > 0$, one may have h_{xX} of either sign. To settle some of these questions we now consider the additional evidence that commissions and margins can furnish.

Equations (13) and (14) give the equilibrium relating the commission t to the marginal benefit and marginal cost and the margin m to the marginal benefit and marginal cost. That is,

$$t = f_x = h_x + g_x h_\sigma \qquad \text{and} \qquad m = f_X = h_X. \tag{24}$$

Hence t and m are functions of α, β, and N. If one can identify the cost or the benefit function using econometric techniques, then this would contribute to our knowledge of the relevant parameters and it would remove some of the indeterminacies above. To this end, we proceed as follows:

$$t_x = f_{xx} = h_{xx} + g_x(2h_{\sigma x} + g_x h_{\sigma\sigma}) + h_\sigma g_{xx}, \tag{25}$$

$$t_X = f_{Xx} = h_{Xx} + g_x h_{\sigma X}, \tag{26}$$

$$t_\alpha = (h_{x\sigma} + h_{\sigma\sigma} g_x) g_\alpha + h_\sigma g_{x\alpha}, \quad \text{(recalling that } \alpha \text{ does not appear in } f\text{)}, \tag{27}$$

$$t_\beta = f_{x\beta} = (h_{x\sigma} + h_{\sigma\sigma} g_x) g_\beta + h_\sigma g_{x\beta}, \tag{28}$$

$$t_N = f_{xN} \quad \text{(recalling that } N \text{ does not appear in } h \text{ or } g \text{ by hypothesis)}. \tag{29}$$

Now one may write

$$dt = t_x\,dx + t_X\,dX + t_\alpha\,d\alpha + t_\beta\,d\beta + t_N\,dN \tag{30}$$

in two different ways by using either the partials from the cost side or the partials from the demand side. Taking a convex combination of the partials from the cost and the demand side gives a continuum of possible ways of writing dt in terms of the underlying structure. That is,

$$\begin{aligned} dt = {} & [\theta(\delta h/\delta x) + (1 - \theta)f_{xx}]\, dx + [\theta(\delta h/\delta X) + (1 - \theta)f_{Xx}]\, dX \\ & + [\theta(\delta h/\delta \alpha) + (1 - \theta)0]\, d\alpha + [\theta(\delta h/\delta \beta) + (1 - \theta)f_{x\beta}]\, d\beta \\ & + [0 + (1 - \theta)f_{xN}]\, dN, \end{aligned} \tag{31}$$

where $0 \leq \theta \leq 1$, and $\delta h/\delta x$, $\delta h/\delta X$, $\delta h/\delta \alpha$, $\delta h/\delta \beta$, and $\delta h/\delta n$ denote the right-hand sides of (25)–(29) where this is applicable. Since the cost function excludes N while the benefit function includes N, a regression of t on the variables on the right-hand side of (31) that excludes dN requires $\theta = 1$, so that the cost function is identifiable.

A similar argument applies to the margin. We have

$$m_x = f_{Xx} = h_{Xx} + h_{X\sigma}g_x, \tag{32}$$

$$m_X = f_{XX} = h_{XX}, \tag{33}$$

$$m_\alpha = h_{X\sigma}g_\alpha \qquad \text{(cf. [27])}, \tag{34}$$

$$m_\beta = f_{X\beta} = h_{X\sigma}g_\beta, \tag{35}$$

$$m_N = f_{XN} \qquad \text{(cf. [29])}. \tag{36}$$

Therefore, a regression of m on the explanatory variables excluding dN would identify the pertinent parameters of the cost function.

This theory determines the equilibrium values of σ together with the equilibrium values of x and X. Given x and X, the distribution of market clearing prices is asymptotically normal with a standard deviation depending on x. The values of x and X give the maximum net benefit. Hence the theory determines the optimal value of σ which takes into account the marginal benefit and the marginal cost of trade on an organized exchange. Consequently, σ is an endogenous variable in this theory.

Section VI continues this analysis empirically.

V. The Sources of Price Variability

Since the empirical evidence refers to physical commodities, not financial instruments such as foreign exchange or bonds which sometimes also have organized futures markets, we parcel out the sources of price variability among the explanatory variables regarded as random variables in the following function:

$$\log p_{it} = a_{0i} + a_{1i} \log q_{it} + a_{2i} \log y_t + u_{it}, \tag{37}$$

where

p_{it} = U.S. price of commodity i in year t during the month of peak stocks divided by the U.S. wholesale price index,[3]

q_{it} = stocks of commodity i in year t during the month of peak stocks in the United States or the world, depending on which is more suitable,

y_t = U.S. gross national income divided by the U.S. consumer price index,

u_{it} = residual.

The coefficient a_{1i} gives the elasticity of the price with respect to the stock, called EPQ; the coefficient a_{2i} gives the elasticity of the price with respect to deflated national income, called EPY; VPAV is the coefficient of variation of the annual average price of the commodity; STDERR is the standard deviation of the residual. We also measure the variability of monthly prices in terms of a parameter called VPM2 (see the Appendix for the details). All of these measures are in percentage terms and are dimensionless numbers comparable across commodities.[4]

In terms of the theory in the preceding section, these statistics are among the determinants of β which enter the benefit function. They are also among the determinants of α which affect σ in (7′). We claim that α and β are increasing functions of EPQ, EPY, VPAV, STDERR, and VPM2.

We take the explanatory variables in (37) as exogenous. First, the more elastic the underlying supply and demand curves for the commodity the closer a_{1i} is to zero. Hence annual variability of stocks has less effect on annual price variability as a_{1i} approaches zero. For example, the better the substitutes for a given commodity in either demand or supply the closer a_{1i} is to zero. Second, price is less responsive to income as a_{2i} approaches zero. The level of income affects the demand for the commodity. As a result, a_{2i} is the compound effect of the size of the income elasticity of demand and the size of the elasticity of supply. The larger is the income elasticity of demand, the larger is a_{2i}. The smaller is the elasticity of supply, the larger is a_{2i}. Therefore, a high income elasticity of demand and a low elasticity of supply combine to give a large value of a_{2i}. The standard deviation of the residuals measures the effects of exogenous factors on price variability that are uncorrelated with the variables explicitly included. Since VPM2 gives the mean value of the standard deviation of prices around the monthly means, it measures the effects of short-term price variability in contrast to the preceding statistics, which give measures of different aspects of annual price variability.

---

[3] Under a system of market-determined foreign exchange rates, the nominal price in U.S. dollars of internationally traded goods, the goods in our sample for the most part, would depend on the factors that determine the U.S. aggregate price level. Under fixed exchange rates this is no longer true. The prices in U.S. dollars of internationally traded goods respond to the forces determining the U.S. aggregate price level only insofar as these forces also influence the world market for these goods. The U.S. economy is a substantial fraction of the world economy. Therefore, even with fixed foreign exchange rates, the prices of internationally traded goods denominated in U.S. dollars depends on the factors influencing the aggregate price level in the U.S. Our analysis treats the U.S. relative price as an endogenous variable. It is, therefore, consistent with floating exchange rates determined by free market forces or with fixed exchange rates and the U.S. economy as a substantial fraction of the world economy. To the extent that the U.S. dollar was a reserve currency during the time of fixed exchange rates, the case for treating the U.S. price level as a function of U.S. monetary and fiscal policy is even stronger. Our empirical experiments with various alternatives led us to the formulation we use.

[4] Even if the R^2 in (37) were 1 for one or even for all commodities, this would not affect the logic of the decomposition of the sources of price variability in terms of the explanatory variables of (37). One may regard (37) as giving an analysis of variance of the price in terms of the variability of the explanatory variables, log q_{it} and log y_t, regarded as random variables. To the extent that log y_t and log q_{it} are predictable 1 or more years in advance it would be desirable to represent this in the analysis of variance. For present purposes this is outside the scope of our analysis.

The sample has 51 commodities for the period 1959-71 (an appendix, available to the reader on request, describes the sources). We divide the sample into three subsamples of equal size such that the first has 17 actively traded goods, the second has 17 less actively traded goods, and the third has 17 with little or no trade as far as we know. Each commodity in a subsample has two companions, one in each of the other two subsamples, with roughly similar gross characteristics—metal, meat, fowl, wood, etc. We use three criteria to place a commodity into a subsample—the size of the open interest, the volume of trade, and the duration and timing of trading activity. Each subsample has commodities difficult to place. For example, we put plywood and broilers in the active group and wool and oats in the inactive group, although the latter two have a larger open interest than the former two. We took as decisive the fact that plywood and broilers have a larger volume of trade. If trade in a commodity came to a halt at some point in the sample period and did not revive during the remainder of the sample period, we put that commodity into the inactive group. This explains our placement of cottonseed meal, coffee, and zinc. A few commodities in the inactive group are actively traded in London, such as tin and zinc. Cocoa is active in both London and New York We put cocoa in the active group. Copper, in the active group, is more active in London than in New York. There have been attempts to start futures markets for some commodities in the dormant group that so far have failed. These commodities are nickel, apples, hams, and rice. We have been unable to discover any attempts to organize futures trading for some commodities in the dormant group, including newsprint, lamb, veal, and frozen strawberries. Except for platinum, our active group includes every commodity actively traded on an organized futures market in the United States during the sample period.

In some cases the absence of an organized futures market probably results from highly stable prices due to governmental policies or actions of private concerns. Foreign exchange, though not in our sample, best illustrates this point. Market determined foreign exchange rates began in 1971. Before the United States closed its gold window the fiscal authorities of various nations had a coordinated price-support program for foreign exchange that maintained a web of fixed exchange rates. Prior to 1971 futures markets in foreign exchange were unimportant. Similarly, the U.S. government price-support program in peanuts and tobacco stifles interest in futures trading in these commodities. Trading in cotton, a commodity with an organized spot market in New York since the eighteenth century, virtually expired when government stocks were large and growing. Trading did not revive until after the disappearance of the government inventory. Published prices of petroleum crude oil and newsprint are very stable. This seems to have the same effect as a price-support program, but we know of no government policy as the explanation. Perhaps the Texas Railroad Commission is to blame for stable crude-oil prices. The end of our sample period precedes the Arabian oil embargo to the United States by 2 years.

Accepting this classification of the commodities into the three groups, we now examine the pertinent evidence. Table 1 gives the means and standard deviations of the variables related to price variability.

According to table 1, VPAV is highest for the active group, next highest for the inactive group, and lowest for the dormant group. This pattern could occur by chance with a probability of 1 percent. Both VPM2 and STDERR show the same pattern. Therefore, prices are most variable for the actively traded group, next most variable for the inactive group, and least variable for the dormant group.

Consider now the pattern with respect to the sources of price variability. The EPQ approaches zero as we move from the active to the dormant group. This is clearer if, instead of calculating the percentage price change with respect to a 1 percent change in stocks, we compute the percentage price change with respect to a 1 percent change of the coefficient of variation of stocks, a parameter denoted QRSK (quantity risk). The inverse relation between trading activity and QRSK is significant at the 99 percent level. Also, EPY decreases with trading activity but the pattern has a low level of significance, 83 percent. Finally, observe that the R^2 (RSQ) and DW statistics do not differ significantly among the three subsamples.

Table 1 shows two other aspects of the difference among commodities that are relevant for our argument. First, the seasonal index of prices is positively correlated with the marginal cost of storage and with the holding of stocks. Observe that the price seasonal varies inversely with trading activity. That is, the actively traded commodities show a more

TABLE 1

MEANS AND STANDARD DEVIATIONS OF SELECTED CHARACTERISTICS RELATED TO PRICE VARIABILITY OF 17 ACTIVE, 17 INACTIVE, AND 17 DORMANT COMMODITIES

Parameter	Active 1–17	Inactive 18–34	Dormant 35–51	Total 1–51	Significance Level*
VPAV	16.4610 (6.6498)	12.7476 (4.7874)	11.5544 (6.0370)	13.5877 (6.1319)	.010
VPM2.	5.5363 (3.0607)	4.7726 (3.9698)	2.9326 (2.3594)	4.4139 (3.3277)	.016†
EPQ	−0.6007 (0.8914)	−0.3628 (0.9100)	−0.2322 (0.4596)	−0.3986 (0.7814)	.071
QRSK	11.6102 (8.9936)	8.2079 (7.1992)	4.8470 (3.5411)	8.2217 (7.3660)	.006
EPY	0.1557 (0.9810)	0.1118 (0.4470)	0.06297 (0.5439)	0.1102 (0.6841)	.170
RSQ.	0.5264 (0.2706)	0.4962 (0.2660)	0.4863 (0.2628)	0.5029 (0.2617)	.176
STDERR.	11.3271 (4.6238)	8.7645 (4.3014)	6.6858 (4.0692)	8.9258 (4.6622)	.004
DW	1.6900 (0.4803)	1.6929 (0.4734)	1.6412 (0.6224)	1.6747 (0.5197)	.199
Seasonal of prices .	29.4229 (35.1045)	22.6903 (19.33)	14.2013 (12.3177)	21.4421 (24.71)	.018†
Seasonal of stocks . .	49.1009 (33.5402)	50.2590 (30.5552)	60.1442 (33.6604)	52.8282 (31.9250)	.090

* The numbers in parentheses are the estimated standard deviations of the population. The statistic as follows

$$[(m_1 - m_2)/\sqrt{(n_1 s_1^2 + n_2 s_2^2)}]\sqrt{[n_1 n_2 (n_1 + n_2 - 2)/(n_1 + n_2)]}$$

has a t-distribution where m_i denotes the sample mean and s_i^2 the sample variance. Let $i = 1$ denote the active, $i = 2$ the inactive, and $i = 3$ the dormant. The theory predicts the signs of $m_2 - m_1$ and $m_3 - m_2$. The significance level gives the probability of observing the two independent t-statistics on the hypothesis that the true differences are zero (Cramèr 1946, sec. 31.2.1).

† There is a slight complication because we lack monthly price series for two commodities, tomato paste, which is in the inactive group, and strawberries, which is in the dormant group. Therefore, we cannot calculate VPM2 and the price seasonal for these two commodities. Consequently, with respect to these two parameters, the sample size is 16 for both the inactive and the dormant group. The significance levels take this into account.

pronounced seasonal price pattern than do the less actively traded ones. The pattern is significant at the 98 percent level. It is also true that there is a more highly seasonal pattern of stocks for the actively traded goods than for the others but with a significance level of only 91 percent. (The stock seasonal is the trough stock level divided by the peak level, where the peak and trough refer to an appropriate 12-month period.) Nevertheless, consistent with our theory that the introduction of futures trading permits a sorting of transactions over time, we find more seasonality for the actively traded goods, which implies a greater benefit from organized futures trading in these goods.

VI. Effects of Price Variability on the Open Interest and the Volume of Trade

The preceding material shows the presence of an inverse association between trading activity and price variability across the three subsamples. According to the theory of the costs and benefits of futures trading given above, the relation between the volume of trade, open interest, and various measures of price variability does not necessarily imply that more price variability raises the volume of trade and open interest. Although greater price variability implies a greater benefit from organized futures trade, it also implies a greater cost. The greater cost is from two sources. First, the higher the price variability, the higher is the cost of holding a given open commitment. Second, the higher the price variability, the larger is the dispersion of the distribution of market clearing prices for a given volume of trade, which is a cost-increasing phenomenon. Therefore, estimates of how the volume of trade and open interest depend on various measures of price variability can give useful information about the nature of the costs and benefits. The purpose of this section is to furnish some empirical evidence in order to narrow the range of theoretical alternatives.

We now use a sample of 25 commodities, which adds data for eight commodities in the inactive sample to the 17 actively traded commodities.[5] The eight additional observations include all commodities in the less active group for which there was some nonnegligible volume of trade and open interest during the sample period. Table 2 lists the commodities and the pertinent data. The velocity of the open interest, the average annual volume of trade divided by the average end-of-quarter open interest, ranges from a low of 14 for coffee to a high of 86 for pork bellies. These are large numbers by comparison with, say, the velocity of the money supply. Soybean future contracts have a daily average volume of trade of $110 millions, and for pork bellies this figure is $66 millions! Table 3 gives the simple correlations, means, and standard deviations of the variables in the regression analysis. Observe that the simple correlations between the log of open interest and price variability are all small, and the same is true of the corresponding correlations for the log of volume of trade. However, since the various measures of price variability are dimensionless pure numbers, while the volume of trade and open interest are in dollars and depend on the size of the demand for the commodities, the two sets of variables are not comparable. Also, the theory implies the presence of a scale variable as a separate factor to explain the size of the open interest and volume of trading in the commodity. The level of stocks suggests itself as one way of measuring scale. This has the

---

[5] The 25 commodities in this sample are as follows: 1-18, 20, 21, 25, 27, 28, 29, and 30 (see table 2 for the names of these commodities).

TABLE 2

DATA BY COMMODITY

No and Commodity	Volume of Trade (10^9 \$)	Open Interest (10^6 \$)	Stocks (10^6 \$)	Seasonal of Stocks	VPAV (%)	STDERR (%)	VPM2 (%)
1. Cotton	1.116	52.6	1374	32.559	16.928	11.82	2.85
2. Soybeans	27.510	552.8	1453	14.127	10.661	5.53	4.42
3. Wheat	9.173	195.8	1418	48.441	15.824	8.10	4.15
4. Cocoa	2.096	124.0	82	20.987	26.902	7.58	7.90
5. Sugar	2.646	110.5	264	42.614	10.875	6.57	5.21
6. Copper	1.073	31.1	244	100.0	23.794	7.72	2.14
7. Silver	5.935	376.3	450	100.0	27.701	11.81	3.88
8. Corn	7.522	216.8	3045	30.354	8.149	5.67	3.49
9. Soymeal	1.808	94.1	497	14.127	12.509	9.58	5.34
10. Soy oil	3.705	95.5	351	14.127	15.304	13.22	6.31
11. Plywood	0.835	16.9	503	96.053	8.614	14.95	6.96
12. Broilers	0.508	10.0	299	75.868	8.348	19.75	2.71
13. Cattle	4.984	170.6	5951	79.178	14.227	6.68	3.28
14. Eggs	1.649	23.8	982	92.738	12.967	19.70	8.24
15. Frozen orange juice	0.715	24.6	181	32.478	25.687	14.21	4.79
16. Pork bellies	16.573	192.5	620	15.448	21.582	14.44	7.54
17. Potatoes	0.659	19.3	416	25.619	19.766	15.24	14.90
18. Wool	0.231	22.1	205	41.190	17.823	16.82	3.38
19. Cottonseed meal	0.002	0.1	149	7.470	11.884	10.85	5.02
20. Rye	0.610	14.2	25	40.079	6.681	3.97	3.37
21. Coffee	0.246	17.8	173	87.735	15.107	13.05	3.63
22. Propane	0.018	3.2	318	48.859	7.390	4.04	1.23
23. Tin	0.003	0.4	72	88.279	20.070	7.19	4.05
24. Zinc	0.011	1.5	45	94.934	11.201	5.17	1.74
25. Oats	0.457	17.2	539	34.598	4.672	4.49	3.06
26. Milo	0.023	2.4	445	41.020	8.184	7.22	3.56
27. Cottonseed oil	0.475	16.0	170	32.426	16.078	8.29	5.95
28. Lumber	0.390	10.1	625	96.053	15.303	8.47	4.07
29. Turkeys	0.162	4.6	384	27.326	9.160	5.88	3.98
30. Hogs	0.374	8.2	1947	83.226	17.770	10.50	5.94
31. Butter	0.056	0.0	114	48.426	6.908	3.10	1.88
32. Tomato paste	0.0	0.0	102	12.114	17.149	10.27	4.77
33. Lard	0.051	71.7	10	65.426	16.321	16.76	6.41
34. Onions	0.0	0.0	14	5.242	15.008	12.91	19.10
35. Tobacco	0.0	0.0	2755	65.026	9.345	3.36	0.83
36. Peanuts	0.0	0.0	282	15.705	9.088	4.82	2.79
37. Rice	0.0	0.0	389	12.629	4.446	4.90	1.74
38. Tea	0.0	0.0	34	60.000	7.549	4.33	1.59
39. Coal	0.0	0.0	394	86.170	25.416	13.42	1.77
40. Aluminum	0.0	0.1	728	100.00	7.620	3.39	1.23
41. Nickel	0.0	0.0	129	100.00	22.544	8.27	2.54
42. Barley	0.0	0.0	412	34.088	8.243	6.58	3.28
43. Petroleum crude oil	0.0	0.0	779	93.637	4.901	1.25	0.66
44. Peanut oil	0.0	0.0	14	15.705	12.173	13.55	4.36
45. Newsprint	0.0	0.0	88	100.00	5.459	1.35	0.56
46. Lamb	0.0	0.0	249	75.00	13.363	8.32	3.86
47. Veal	0.0	0.0	278	79.178	12.571	6.39	5.34
48. Cheese	0.0	0.0	158	74.445	19.852	4.27	1.93
49. Strawberries	0.0	0.0	19	41.806	11.850	10.10	2.93
50. Hams	0.0	0.0	1057	67.833	8.701	5.22	4.08
51. Apples	0.0	0.0	197	1.230	13.304	14.13	10.35

NOTE.—A detailed description of the sources and of the methods for estimating these figures is available to readers on request.

TABLE 3

SIMPLE CORRELATIONS BETWEEN LOG VOLUME, LOG OPEN, AND SELECTED MEASURES OF PRICE VARIABILITY, MEANS, AND STANDARD DEVIATIONS FOR 25 COMMODITIES

VARIABLES	SIMPLE CORRELATIONS					MEAN	SD
	Log Open	Log Volume	VPM2	STDERR	VPAV		
Log open	1	.93267	−.02780	−.21903	.25128	3.75921	1.29777
Log volume	...	1	.05703	−.17166	.13909	0.31176	1.36342
VPM2	..	...	1	.35044	.24315	5.09960	2.60072
STDERR	...	...	...	1	.24568	10.56160	4.51223
VPAV	..	. .	...	...	1	15.29728	6.22404

disadvantage of introducing serious measurement errors for some commodities. As an alternative, one can use the volume of trade as a scale variable to explain the open interest and the open interest as a scale variable to explain the volume of trade. In effect this treats the system of equations (15) as candidates for direct estimation by least squares.

Substitute β for z in (15) and write out the result as follows:

$$J_{xx}(dx/d\beta) + J_{Xx}(dX/d\beta) = -J_{x\beta}, \tag{38}$$

$$J_{Xx}(dx/d\beta) + J_{XX}(dX/d\beta) = -J_{X\beta}. \tag{39}$$

Take $dx/d\beta$ as the dependent variable in the first and $dX/d\beta$ as the dependent variable in the second to give the following representation:

$$dx = -(J_{Xx}/J_{xx})\, dX - (J_{x\beta}/J_{xx})\, d\beta, \tag{40}$$

$$dX = -(J_{Xx}/J_{XX})\, dx - (J_{X\beta}/J_{XX})\, d\beta. \tag{41}$$

Take a linear approximation to these equations and use the three measures of price variability for β. We estimate by least squares the following equations:

$$\begin{aligned} \log x = {} & a_0 + a_1 \log X + a_2(\text{VPM2}) + a_3(\text{STDERR}) \\ & + a_4(\text{VPAV}) + \varepsilon, \end{aligned} \tag{42}$$

$$\begin{aligned} \log X = {} & c_0 + c_1 \log x + c_2(\text{VPM2}) + c_3(\text{STDERR}) \\ & + c_4(\text{VPAV}) + \zeta. \end{aligned} \tag{43}$$

There is the complication that x and X are endogenous variables. It seems plausible that the random disturbances are more likely to shift the benefit functions than the cost functions, because changes in weather affect production of agricultural goods and cyclical factors affect the demand for all goods, which shifts the hedging demand and, therefore,

TABLE 4

SELECTED STATISTICS FROM REGRESSIONS RELATING FUTURES TRADING ACTIVITY TO PRICE VARIABILITY FOR A SAMPLE OF 25 COMMODITY FUTURES

	DEPENDENT VARIABLES			
	Log Open Interest		Log Volume of Trade	
EXPLANATORY VARIABLES	Elasticity	t-Ratio of Coefficient	Elasticity	t-Ratio of Coefficient
Log volume	0.8591	12 412	...	...
Log open	..	..	1.0302	12.412
VPM2	−0.2406	−1.240	0.2793	1.321
STDERR	−0.2182	−0.920	0.1329	0.504
VPAV	0.1769	2.235	−0.4790	−1.773
R^2	.9020	. .	.8953	...

the benefit function. Therefore, ε will be positively correlated with log X and ζ will be positively correlated with log x. Hence least-squares estimates of the coefficients, a_1 and c_1, are both upward biased. We deal directly with one aspect of this problem below by studying how commissions and margins depend on the pertinent variables as given in (24).

Table 4 gives the estimates of the elasticities derived from (42) and (43) together with the t-ratios of the coefficients. The coefficient of log X in (42) corresponds to $-(J_{Xx}/J_{xx})$ in (40) while the coefficient of log x in (43) corresponds to $-(J_{Xx}/J_{XX})$ in (41). The other coefficients in (42) and (43) relate similarly to their counterparts in (40) and (41). Consider the regression with the log of the open interest as the dependent variable. Of the three measures of price variability, the only one that enters with a large t-ratio is VPAV, the coefficient of variation of annual prices. It has a t-ratio of 2.2 and an elasticity of 0.18. The other two measures of price variability enter with t-ratios around 1 and have negative elasticities. Also observe that in the regression equation with log x as the dependent variable each coefficient of price variability enters with a sign opposite to that which it has in the regression with log X as the dependent variable. One is immediately tempted to conclude from these results that a rise in some aspect of price variability tends to move the open interest and the volume of trade in opposite directions. This may be true, but the high correlation between log x and log X is consistent with an alternative and a simpler explanation of these results. The simple correlation between log X and log x is 0.933 according to table 3. Also, the multiple R^2 in the log open regression in table 4 is .902 (R = .95). One may then say that these regressions are nearly identities and it would follow by simple arithmetic that the coefficients of price variability in one equation are nearly perfectly predictable by the coefficients in the other. For instance, solve for log x in the regression equation with log X

TABLE 5

SELECTED STATISTICS FROM REGRESSIONS TO EXPLAIN LOG OF OPEN INTEREST AND LOG OF VOLUME OF TRADE FOR 25 COMMODITY FUTURES

Explanatory Variables	Dependent Variables			
	Log Open Interest		Log Volume	
	Elasticity	*t*-Ratio of Coefficient	Elasticity	*t*-Ratio of Coefficient
Log stock	0.4451	2.13	0.5594	2.47
VPM2	0.0167	0.03	0.3010	0.55
STDERR	−0.9155	−1.49	−0.8122	−1.24
VPAV	1.2109	1.93	0.8031	1.20
R^2	.3048	...	.2893	...

as the dependent variable and you will come close to the estimated regression with log x as the dependent variable.

For these reasons it is of some interest to try an alternative to these equations which uses the log of stocks as a scale variable. These results appear in table 5. Here we see that each measure of price variability has a coefficient of the same sign in the regressions with log X and log x as the dependent variables. This is hardly surprising given the high positive correlation between log X and log x. Nevertheless it is worth noting that an increase in stocks of 1 percent will lead to a larger percentage increase in the volume of trade than in the open interest.[6]

We now study how the commission per transaction and the margin per contract depends on the pertinent variables according to the theory in Section IV.[7] We assume that the cost and benefit functions are of the same form for all commodities and that differences among commodities express themselves via different values of the exogenous variables, α, β, and N. In equilibrium the commission, t, and the margin, m, satisfy (24). The argument in Section IV shows that the cost functions are identified in a regression that excludes the scale variable N and includes the other exogenous variables related to α and β. We have two regressions that we estimate by ordinary least squares. In the first the commission is the dependent variable and in the second the margin is the dependent variable. In addition to the proxies for α and β, we include among the explanatory variables, $\log x$, $\log X$, and the average price per futures contract over the sample period, a variable we label CONPR (contract price). The relevance of CONPR as a determinant of m hardly needs discussion. Plainly, the potential cost to a broker of a customer default is an increasing function of the size of the customer's commitment or open interest. Hence m should be an increasing function of CONPR. It is not equally plain why CONPR should affect the commission. Empirically, one can hardly doubt that t is an increasing function of CONPR. This result is consistent with the belief that it is more costly to arrange a mutually satisfactory trade for a contract of larger value, such as coffee, than for one of lesser value, such as potatoes, even holding the volume of trade constant. That is, a broker's execution according to his instructions is more costly for a contract of a larger value. It may not be superfluous to add that this argument would be consistent with the theory of optimal search.

---

[6] It has not escaped our attention that one may use the regression equations in table 5 to predict the log of the open interest and the log of the volume of trade for the 26 (= 51 − 25) commodities which now lack an active organized futures market. In our earlier work we did precisely this. We have doubts about the reliability of these results for several reasons. The most important caveat is with respect to the measure of stocks to represent the scale of potential interest in the commodity. Empirically, it turns out to be a very important variable for explaining lack of interest in a commodity. That is, commodities with a small potential market are bad candidates for an organized futures market. In some cases there are large government holdings of the commodity (e.g., peanuts and tobacco), and it is difficult to determine how to measure the relevant stocks. (For empirical work on this problem, see Telser 1958.) There may also be a price-stabilizing effect of organized futures trading so that the relevant measure of price variability between traded and untraded goods is larger than would appear from the observed differences in price variability. (For a theoretical analysis of the stabilizing effects on prices of profitable speculation, see Telser [1959].) A third important factor, the tax incentive for trading in futures contracts in the United States, does not appear in our theoretical or empirical analysis. Such trading may allow a shift of capital gains and losses from various speculations between years and thereby lowers the income tax of the speculator, who can take advantage of the lower tax rate on capital gains. The trade believes this is important. (For a careful analysis of portfolio aspects of futures trading, see Dusak [1973].)

If a good measure of scale were available, this would enable the application of simultaneous-equations techniques in an attempt to estimate the structural relations of the model. The only candidate is an estimate of stocks by commodity, and this is subject to considerable measurement error. Therefore, we focus our attention on the cost side which is identifiable and we accept a possible simultaneous equations bias in so far as $\log x$ and $\log X$ are both positively correlated with the variable we omit, the scale factor. This gives an upward bias to the coefficients of both $\log x$ and $\log X$ if t is positively related to N and m is positively related to N, where N denotes the scale factor.

---

[7] We use a sample of 23 commodities. The two missing commodities are cottonseed oil and turkeys. Both have been inactive since 1965. We lack data giving the commissions and margins for these two, and this explains their absence from our sample. The commission per transaction and the margin per contract were obtained from brokerage firms and refer to the first half of 1976. The sample period gives averages of the explanatory variables for an earlier period. This should cause no problem. One may argue that the commissions and margins do in fact depend on long-term averages. In any case, since these variables appear as the dependent variables in the regression equations, the presence of measurement error does not bias the estimates of the regression coefficients. However, it does lower the R^2 and increase the standard errors of the estimated regression coefficients.

TABLE 6

SELECTED STATISTICS FROM REGRESSIONS TO EXPLAIN COMMISSIONS AND MARGINS FOR A SAMPLE OF 23 COMMODITY FUTURES

EXPLANATORY VARIABLES	DEPENDENT VARIABLES			
	Commission		Margin	
	Elasticity	t-Ratio of Coefficient	Elasticity	t-Ratio of Coefficient
Log volume	−0.218	−2.72	−0.419	−2.80
Log open interest	0.141	1.58	0.449	2.69
CONPR	0.536	4.62	0.729	3.36
STDERR	−0.189	−1.75	−0.425	−2.10
VPAV	0.076	0.72	−0.265	−1.33
VPM2	0.130	1.48	0.415	2.53
Constant	−0.1997	−0.01	−1,227.00	−1.59
R^2	.688	...	.662	...
SE	8.678	...	397.1	...

We claim that the regressions given in table 6 using the data in table 7 furnish estimates of the slopes of the commission and margin with respect to the various variables for the cost equation. Thus, the slope of t with respect to x in table 6 gives an estimate of t_x on the cost side, which is the right-hand side of (25), and so on. Similarly, the slope of m with respect to x gives an estimate of $h_{Xx} + h_{X\sigma}g_x$ according to the right-hand side of (32), and so on. With the theoretical interpretations in hand as given by (25)–(29) and (32)–(35), consider the evidence in table 6.

Both the commission and the margin regressions display a similar pattern of results.[8] With only one exception all of the explanatory variables have coefficients of the same sign in both regression equations. The higher is the volume of trade the lower is the commission and the margin. The higher is the open interest, the higher is the commission and the margin. The higher is the average price per contract, the higher is the commission and margin. Although the t-ratios indicate a high level of significance by the usual standards, one should regard this with some skepticism because, for one thing, simultaneous-equations bias is present due to the omission of a scale variable that may be positively correlated with $\log x$ and $\log X$. The coefficients of $\log x$ and $\log X$ both give strong support to the belief that $g_x < 0$ (cf. [25] and [32]). The negative coefficient of $\log x$ in the margin regression is also consistent with the view that a rise in the volume of trade for a given open interest lowers the marginal cost with respect to the open interest ($h_{Xx} < 0$).

Now consider the regression coefficients of the various measures of price variability in table 6. In the commission regressions these coefficients

---

[8] The simple correlation between m and t is 0.643. This seems too low to explain the similarity in the pattern of results.

TABLE 7

AVERAGE CONTRACT PRICE (CONPR), COMMISSION (COMM), AND MARGIN (MARG) FOR 23 COMMODITY FUTURES CONTRACTS IN DOLLARS PER CONTRACT

Commodity	CONPR	COMM	MARG
Cotton	14,368	70	1,300
Soybeans	13,715	50	2,500
Wheat	8,858	45	1,250
Cocoa	8,511	70	2,400
Sugar	8,034	75	2,000
Copper	9,949	50	750
Silver	7,926	50	1,200
Corn	6,100	40	1,000
Soy meal	7,527	40	1,500
Soy oil	6,005	40	1,000
Plywood	5,947	40	700
Broilers	6,250	35	500
Cattle	10,925	50	900
Eggs	8,071	40	700
Frozen orange juice conc.	6,712	45	450
Pork bellies	14,425	60	1,500
Potatoes	1,892	35	750
Wool	6,947	50	1,500
Rye	5,941	40	1,500
Coffee	12,764	80	2,000
Oats	3,357	35	750
Lumber	4,198	55	700
Hogs	4,087	40	900

refer to t_α or t_β in (27) and (28). If $h_{x\sigma} + h_{\sigma\sigma}g_x$ is negative because g_x is negative, then the coefficient of the measure of price variability can also be negative. Nor is this all. Given the hypothesis about the signs of the various partials in (27) and (28), a negative value of g_x is *necessary* but is not sufficient for an observed negative value of the coefficient of a price variability parameter; STDERR does have a negative coefficient in the commission regression. Both VPAV and VPM2 have positive coefficients in the commission regression.

Next consider the margin regression. Equations (34) and (35) predict positive coefficients of the variables that are positively related to price variability if $h_{X\sigma}$ is positive and if X is not an argument in $g(\cdot)$. According to table 6, VPM2 does have a positive coefficient with an elasticity of 0.415 and a t-ratio of 2.5. Recall that VPM2 measures the coefficient of variation around the monthly price seasonal so that it reflects shorter term sources of price variability than either VPAV or STDERR. Both STDERR and VPAV have negative coefficients. The elasticities are -0.425 and -0.265, and the t-ratios are -2.1 and -1.3. These results are not consistent with the present model. There are two possible explanations of this discrepancy of which the first is econometric and the second is economic. The regression equation omits the scale variable, N. According to the theory N is positively related to x and X and is also positively related to m (and to t) as is shown in (36) (and [29]). This imparts a positive bias to the estimates of the coefficients of log x and log X. It is also possible that N is correlated with the exogenous variables α and β. If N is negatively related to α and β, then this would impart a downward bias to the estimates of the coefficients of price variability. We do have a weak proxy for the scale variable given by the log of stocks, a variable which we regard with suspicion because of measurement error for some commodities. Nevertheless, we do find negative, albeit small, correlations between the three measures of price variability and the log of stocks. The simple correlations, denoted by $R(\cdot)$, are as follows:

$$R(\text{log stock, VPAV}) = -.149, \qquad R(\text{log stock, STDERR}) = -.064,$$

$$R(\text{log stock, VPM2}) = -.045.$$

Observe that it is the variable with the correlation closest to zero that has a positive effect on the margin.

This econometric argument agrees with an economic argument. The size of the elasticity of the excess demand function increases with the stock of outstanding commitments. According to (4), this tends to lower σ. We believe that a closer investigation of the function $\sigma = g(x, \alpha, \beta)$ is desirable since these arguments suggest that the variable X should enter this function. Such research would directly study the determinants of liquidity in the market.

VII. Summary and Conclusions

In an organized market the participants trade a standardized contract such that each unit of the contract is a perfect substitute for any other unit. The identities of the parties in any mutually agreeable transaction do not affect the terms of exchange. The organized market itself or some other institution deliberately creates a homogeneous good that can be traded anonymously by the participants or their agents.

Although the discussion centers on organized futures markets, the basic theory applies equally well to any organized market. The benefit of an organized market is an increasing function of the number of potential participants. It is also an increasing function of the turnover of the potential participants in that market. It would not be necessary for a small group of traders who know each other well and who have had and will continue to have contacts with each other to bear the expense of organizing a formal market. In such markets the terms of sale often do depend on the identity of the parties in addition to the characteristics of the goods. An organized market deals in a highly fungible good that is readily traded among strangers. In an organized market the transactions prices alone convey a considerable amount of useful information to those who are not currently trading in the market. In those markets where heterogeneous goods are traded and where the identity of the buyer and seller affects the terms of trade, the transaction price alone conveys only partial information to outsiders.

In addition to scale, price variability affects the benefit of having an organized market. It also affects the cost. There is more price variability for those goods that have an organized futures market than for the goods that lack such markets. It does not follow that futures trading causes greater price variability. The *organization of a futures market is the response* to an increase of price variability. For example, when the government allows the price of a good to fluctuate and abandons its attempt to control the price, this may create an incentive to organize a futures market if the potential scale of operation is large enough, as witness the creation of organized futures markets in some foreign exchange.

We find that the volume of trade increases relative to the open interest, the higher is the level of the open interest. As the open interest in a commodity declines, the volume of trade declines even more rapidly. According to the empirical results relating commissions to log x and log X, commissions are higher, the less active the trade in the commodity. Let price variability be the driving force as represented by the parameter β. Let the price of some good become more stable. The open interest declines and the volume of trade falls relative to the open interest. This raises the commission and the margin. A sequence of events now begins that may well end in a corner equilibrium with no trade and no open interest in the commodity. In this way the theory predicts the disappearance of an organized market. Similarly, a rise in price variability may lead to the appearance of an organized market.

The best concise summary of the theory is as follows: An organized market facilitates trade among strangers.

Appendix

Description of Cross-Section Data and Sources

I. Trading activity and market structure variables

Seasonal of stocks. Average ratio of trough-month stocks to peak-month stocks, 1959–71, expressed as a percentage.

Open. Average number of open contracts, end-of-quarter, 1959–71 or subperiod of active futures trading. *Journal of Commerce and Commercial.*

Opnint. Open × average dollar value of one contract, 1959–71 or subperiod of active futures trading, in millions of nominal dollars.

Stocks. Value of average privately held stocks in millions of nominal dollars, 1959–71 or subperiod of active futures trading. Computations are described in appendix II, Average U.S. Stocks. Conceptually, stocks are computed as pipeline stocks plus average intraseasonal inventories. They give the peak holdings during the year.

Volume. Average annual volume of futures transactions on all U.S. contract markets, 1959–71 or subperiod of active futures trading, × average dollar value of one contract, expressed in billions of nominal dollars. Association of Commodity Exchange Firms, Inc.

II. Variables derived from monthly price series P_{tm}

VPAV. Coefficient of variation across years of $\sum_{m=1}^{12} P_{tm}$, 1959–71. When weighted average yearly prices are available (i.e., weighted by quantities sold), such data are used.

VPM2. Let P_{tm}^{*} equal P_{tm} deflated by a 12-month moving average, centered at the sixth month. For each month m, let σ_m equal 100 × standard deviation of P_{tm}^{*}; VPM2 equals the mean of σ_m across months.

Seasonal of prices. Maximum own rate of interest, at annual rate, displayed by averages across years of P_{tm} (one for each month), 1959–71. Example: if $P\cdot_{\text{DEC}}$ and $P\cdot_{\text{SEP}}$ represent average December and September prices, and if $P\cdot_{\text{DEC}} > P\cdot_{\text{NOV}} > P\cdot_{\text{OCT}} > P\cdot_{\text{SEP}}$, then season equals 400 log $(P\cdot_{\text{DEC}}/P\cdot_{\text{SEP}})$, if $P\cdot_{\text{DEC}}$ and $P\cdot_{\text{SEP}}$ are the maximizing pair. Months are indicated in table 1.

III. Variables from regressions of price on stocks and income

For each commodity, regress via ordinary least squares:

$$\log \left(\frac{\text{PSK}}{\text{WPI}}\right)_t = a_0 + a_1 \log (\text{QSK})_t + a_2 \log \left(\frac{\text{GNI}}{\text{CPI}}\right)_t + u_t, \text{ 1959–71},$$

where

QSK = peak-month stocks
PSK = U.S. nominal spot (monthly) price
GNI = U.S. gross national income. *Statistical Abstract*
WPI = U.S. wholesale price index. *Statistical Abstract*
CPI = U.S. consumer price index. *Statistical Abstract.*

DW. Durbin-Watson coefficient.

EPQ. $\hat{a}_1$.

EPY. $\hat{a}_2$.

ERR. Standard error of regression × 100 (i.e., expressed as a percentage coefficient of variation).

QRSK. $|\hat{a}_1|$ × coefficient of variation across years of QSK.

RSQ. Coefficient of determination (R^2).

STDERR.[9] ERR × $[\text{DW}(1 - 0.25\text{DW})]^{1/2}$.

---

[9] The indicated adjustment is given as follows:

Let

$$u_t = v_t + \rho u_{t-1}, Eu_t v_\tau = 0, t \neq \tau, Ev_t^2 = \sigma^2.$$

Then

$$\text{var } u_t = E(v_t^2 + 2\rho v_t u_{t-1} + \rho^2 u_{t-1}^2) - \rho^2 (Eu_{t-1})^2,$$

$$\text{var } u_t = \sigma^2/(1 - \rho^2). \tag{A1}$$

Since

$$\text{DW} \equiv \sum (u_{t+1} - u_t)^2 / \sum u_t^2,$$

$$E(\text{DW}) = 2[1 - E(\sum u_{t+1} u_t / \sum u_t^2)].$$

Hence

$$E(\text{DW}) = 2(1 - \rho). \tag{A2}$$

Substitution of (A2) into (A1) implies the indicated formula.

References

Arrow, Kenneth J. *Essays in the Theory of Risk Bearing.* Chicago: Markham, 1971.

———. "Vertical Integration and Communication." *Bell J. Econ.* 6 (Spring 1975): 173–83.

Bresciani-Turroni, C. "L'Influence de la speculation sur les fluctuations des prix du coton." *L'Egypte contemporaine* 22 (March 1931): 308–42.

Cramèr, Harald. *Mathematical Methods of Statistics.* Princeton, N.J.: Princeton Univ. Press, 1946.

Demsetz, Harold. "The Cost of Transacting." *Q.J.E.* 82 (February 1968): 33–53.

Dusak, Katherine. "Futures Trading and Investor Returns: An Investigation of Commodity Market Risk Premiums." *J.P.E.* 81, no. 6 (November/December 1973): 1387–1406.

Higinbotham, Harlow N. "The Demand for Hedging in Grain Futures Markets." Ph.D. dissertation, Univ. Chicago, 1976.

Houthakker, H. S. "The Scope and Limits of Futures Trading." In *The Allocation of Economic Resources*, edited by Moses Abromovitz et al. Stanford, Calif.: Stanford Univ. Press, 1959.

Marshall, Alfred. *Industry and Trade.* 3d ed. London: Macmillan, 1920.

Telser, Lester G. "Safety First and Hedging." *Rev. Econ. Studies* 23 (December 1955): 1–16.

———. "Futures Trading and the Storage of Cotton and Wheat." *J.P.E.* 66, no. 3 (June 1958): 233–55.

———. "A Theory of Speculation Relating Profitability and Stability." *Rev. Econ. and Statis.* 41 (August 1959): 295–301.

———. "The Supply of Speculative Services in Wheat, Corn, and Soybeans." *Food Res. Inst. Studies* 7, suppl. (1967): 131–76.

———. *Economic Theory and the Core.* Chicago: Univ. Chicago Press, forthcoming.

Working, Holbrook. "Theory of the Inverse Carrying Charge in Futures Markets." *J. Farm Econ.* 30 (February 1948): 1–28.

———. "Futures Trading and Hedging." *A.E.R.* 43 (June 1953): 314–43.

———. "New Concepts concerning Futures Markets and Prices." *A.E.R.* 52 (June 1962): 431–59.

———. "Toward a Theory concerning Floor Trading on Commodity Exchanges." *Food Res. Inst. Studies* 7, suppl. (1967): 5–48.

Public Property and the Effects of Migration upon Other Residents of the Migrants' Countries of Origin and Destination

---

Dan Usher

Queen's University, Kingston, Ontario

When a man migrates from one country to another, he abandons his share of public property—the use of roads and schools, the rights to a share of revenue from minerals in the public domain, and so on—in the former country and acquires a share of public property in the latter, conferring a benefit upon the remaining residents of the country from which he comes and imposing a cost upon the original residents of the country to which he goes. This proposition is developed with the aid of an ordinary marginal product of labor curve, the "public property" effect is contrasted with other sources of costs and benefits of migration, and an attempt is made to supply a rough and ready estimate of the magnitudes involved.

Who gains and who loses when a man migrates from one country to another? Economists have tended to analyze this question in terms of a standard case with a list of exceptions. The standard case is that of the migrant who was paid the value of his marginal product in his country of origin and who will be paid the value of his marginal product in his country of destination, so that, while he himself may gain because his marginal product is greater in the latter country, his migration has no effect whatsoever on the rest of the population in either country.[1] The exceptions are circumstances when this conclusion need be modified

---

In preparing this paper, I have drawn upon advice and suggestions of my colleague Oli Hawrylyshyn, John Martin and John Flemming of Nuffield College, and Carlos Diaz-Alejandro of Yale. I am grateful to Bram Cadsby, a graduate student at Queen's, for spotting and correcting an error in an earlier draft.

[1] This argument is set out in Grubel and Scott (1966). Specifically the argument is that the flow of scientists and other skilled people from Canada to the United States cannot be harmful to Canada if these people would have received their marginal products had they remained in Canada. A similar view is expressed in papers of Harry Johnson (1965, 1967). Johnson has considered several of the "exceptions" discussed below including the argument about public support of education.

[*Journal of Political Economy*, 1977, vol. 85, no. 5]
© 1977 by The University of Chicago. All rights reserved.

because migration affects factor prices or because the migrant's earnings differ from the value of his marginal product.

The purpose of this essay is to add an item to the list of exceptions. Briefly, the point I wish to make is this: An assessment of the costs and benefits of migration need take account of the fact that a large portion of a country's property is publicly owned, so that a migrant on going from one country to another must as a rule abandon his share of publicly owned property in his country of origin and acquire a share of publicly owned property in his country of destination. The migrant exchanges his right to send his children to school in his country of origin for the right to send his children to school in his country of destination, reducing the need for new school building in the former country and increasing it accordingly in the latter. The migrant on entering a country and on becoming a citizen exchanges rights to the ownership of public buildings of all kinds, of roads, parks, all land in the public domain, royalties on minerals, proceeds from the corporation income tax, and so on with a corresponding deduction for his share of public debt. The effect of this sharing of public property looked upon in isolation from other exceptions to be discussed below is that migration confers a benefit upon the residents of his country of origin and a cost upon the residents of his country of destination. Though analytically simple, even trivial, our proposition is nonetheless important because it seems to stand in direct opposition to the widely held and strongly held view of which the following quotation is representative: Hamada (1977, p. 20) speaks of a "basic reason why moral concerns are raised in the case of brain drain: in contrast to the case of foreign investment where the gain from the international factor movement is divided by the two countries, the DC (developed country) gains now at the cost of those left behind in the LDC (less developed country). The emigrants similarly are seen to gain at the sacrifice of those left behind in the LDC." This view of the effects of migration could stand as the foundation for the belief that people in developed countries "ought" to compensate people in less developed countries for the ill effects of the brain drain.[2] Toward the end of this essay, we consider how the usual justification for this statement stands up when allowance is made for the existence of publicly owned property, and we conclude that, though the statement may still be true on occasion, it could easily be reversed and would be reversed unless migrants are very much richer or more skilled than the average resident in their country of destination. We begin this essay with a statement of the standard case and some well-known exceptions. Then we consider the effects of publicly owned property in theory and in a numerical example. We conclude with a discussion of some qualifications to the main argument.

---

[2] See also Bhagwati et al. (1975), Bhagwati and Partington (1976), and Bhagwati (1976). The case for taxing the migrants themselves does not require either that immigration of educated people is harmful to the remaining residents of the migrants' countries of origin or that migration is beneficial to the original residents of their countries of destination. It is sufficient that the migrant himself benefits and this benefit is a potential source of taxation.

The Standard Case and Some Exceptions

In the "standard case," a migrant has no effect either upon the welfare of the people who remain behind in the country from which he comes or upon the welfare of the original residents of the country to which he goes because he would earn the value of his marginal product in each case. The standard case may be conveniently illustrated in a one-sector model where a uniform output is produced with two factors of production. We shall concentrate upon the costs and benefits of migration to residents of the country where the migrant goes, but the analysis of costs and benefits to his country of origin can be carried out on similar lines.

Consider an economy in which output Q is produced with factors of production, labor and capital, L and K, in accordance with a linear homogeneous production function $Q = f(K, L)$. We are supposing that perfect competition prevails throughout the economy so that the real wage equals the marginal product of labor, $w = \delta Q/\delta L$, and the real rent on capital equals the marginal product of capital, $r = \delta Q/\delta K$. The term "real" in the preceding sentence signifies that w and r are measured in units of Q. From the assumption about the form of the production function, it follows that

$$Q = wL + rK. \tag{1}$$

A migrant may be thought of as bringing with him a bundle of labor and capital, ΔL and ΔK. His presence causes the wage rate to change from w to $w + \Delta w$, the rent on capital to change from r to $r + \Delta r$, and total output to increase from Q to $Q + \Delta Q$. Consequently

$$\begin{aligned} Q &+ \Delta Q \\ &= (w + \Delta w)(L + \Delta L) + (r + \Delta r)(K + \Delta K) \\ &= (wL + rK) + (L\,\Delta w + K\,\Delta r) + [(w + \Delta w)\,\Delta L + (r + \Delta r)\,\Delta K]. \end{aligned} \tag{2}$$

The three expressions on the right-hand side of the equation are the remuneration of the original residents prior to the migration, the change in their remuneration after the migration, and the remuneration of the *migrant himself.* By expanding $Q + \Delta Q$ in a Taylor series and ignoring third and higher terms in the series, we see that[3]

$$L\,\Delta w + K\,\Delta r = -\tfrac{1}{2}(\Delta L\,\Delta w + \Delta K\,\Delta r) \geq 0. \tag{3}$$

This expression is necessarily nonnegative because of the assumed convexity of the isoquants; see Ferguson (1969, p. 56).

What we are calling the standard case is where ΔL and Δr are small by comparison with L and r. In that case, the second expression in equation (2) is virtually zero because each term on the right-hand side of equation (3) is the product of increments. It follows that

$$\Delta Q = (w + \Delta w)\,\Delta L + (r + \Delta r)\,\Delta K \tag{4}$$

and

$$Q = (w + \Delta w)L + (r + \Delta r)K. \tag{5}$$

Equation (4) may be interpreted as the proposition that migrants earn the values of the marginal products of their factors, while equation (5) may be interpreted as the proposition that original residents as a whole are neither better off nor worse off than they were before.

These propositions are illustrated in the marginal product of labor curve in figure 1 for the special case where the migrant has labor but no capital, where $\Delta K = 0$ and $\Delta L > 0$. The value of the national income prior to migration is $LW + Kr$ – the area under the curve to the left of L. The income of the migrant, the second term in equation (2), is the box $BDEC$. The increase in the income of the original residents resulting from migration is the little triangle ABC which becomes infinitely small as a proportion of $BDEC$ when $\Delta L/L$ approaches zero. Ignoring the triangle ABC, we see that *immigration of ΔL increases the national income by* precisely the remuneration of the migrants. It is this proposition that we take as a starting point for examining the "exceptions" that might cause migration to affect the welfare of the original residents favorably and unfavorably as the case may be.

There would seem to be three main classes of exceptions which might be called the terms of trade argument, the distributional argument, and the factor market argument. These broad exceptions can in turn be subclassified. We shall consider the exceptions briefly one by one.

---

[3] Expanding $f(L + \Delta L, K + \Delta K)$ and setting $f_l = w$ and $f_k = r$, we see that

$$\begin{aligned}
Q + \Delta Q &= f(L + \Delta L, K + \Delta K)\\
&= f(L, K) + f_L\,\Delta L + f_K\,\Delta K + \tfrac{1}{2}(f_{LL}\,\Delta L^2 + f_{KK}\,\Delta K^2 + 2f_{LK}\,\Delta L\,\Delta K)\\
&= wL + rK + w\,\Delta L + r\,\Delta K + \tfrac{1}{2}\left(\frac{\partial w}{\partial L}\,\Delta L + \frac{\partial w}{\partial K}\,\Delta K\right)\Delta L\\
&\quad + \tfrac{1}{2}\left(\frac{\partial r}{\partial L}\,\Delta L + \frac{\partial r}{\partial K}\,\Delta K\right)\Delta K\\
&= w(L + \Delta L) + r(K + \Delta K) + \tfrac{1}{2}\Delta w\,\Delta L + \tfrac{1}{2}\Delta r\,\Delta K\\
&= Q + \Delta Q - (L + \tfrac{1}{2}\Delta L)\,\Delta w - (K + \tfrac{1}{2}\Delta K)\,\Delta r.
\end{aligned}$$

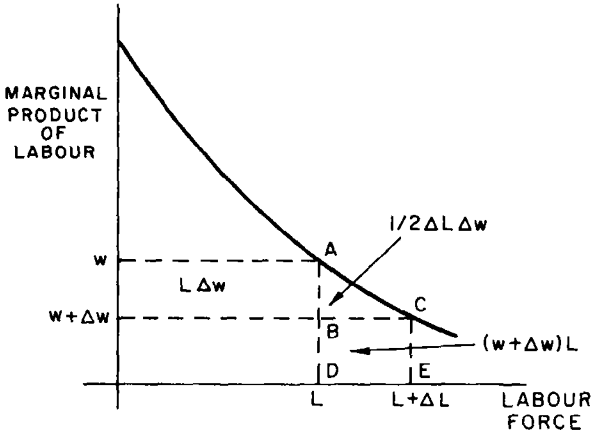


FIG. 1

The Terms of Trade Argument

The general case.—The original residents as a group will benefit from migration which is on a large enough scale and where migrants are sufficiently different from the original residents in their endowments of labor and capital to affect the wage-rental ratio significantly. To see this point, suppose once again that migrants have no capital, $\Delta K = 0$, but now let ΔL be a finite proportion of L so that the wage rate falls by a measurable amount, Δw. From equation (3), it follows at once that the gain to old residents as a group is

$$L\,\Delta w + K\,\Delta r = -\tfrac{1}{2}\,\Delta L\,\Delta w > 0. \tag{6}$$

The gain to the original residents is illustrated in figure 1 by the area of the triangle ABC, to which $\frac{1}{2}\,\Delta L\,\Delta w$ in equation (6) is a close approximation. As the labor force increases from L to $L + \Delta L$, the national income increases by the area $ADEC$ under the marginal product of labor curve. Of that increase, only $BCED$ is paid to the migrants themselves, leaving the triangle ABC for the original residents. The remuneration of the labor of the original residents is reduced by $L\,\Delta w$ but that is not a social loss because it is made up by an equal and opposite change in the remuneration to capital. The net benefit to the old residents arises entirely from the reduction in the remuneration of the "earlier" migrants resulting from the appearance of the "later" ones. This source of gain from migration is a variant of the normal gain from trade. A group of people owning factors L_1 and K_1 are made better off if they are permitted to trade with a second group of people owning factors L_2 and K_2, as long as $L_1/K_1 \neq L_2/K_2$. This follows immediately from equation (3) by setting $L_1 = L$,

$L_2 = \Delta L$, and so on. The argument summarized in this paragraph is well developed in Berry and Soligo (1969); it is also contained in Borts and Stein (1964, p. 192).

Migrant-specific factors of production.—Though the general case above has received a good deal of attention, there is some question about its quantitative importance. Does migration really make a country better off because the capital-labor ratio of migrants is different from the capital-labor ratio of original residents? One cannot help suspecting that the little triangle ABC in figure 1 will, like many of the little triangles in economics, prove to be uninterestingly small.

On the other hand, the analysis looks quite different if we suppose the migrant to be the bearer of a unique factor of production and if we designate the original residents to be the income recipients from all other factors of production. The Chinese restaurant is a case in point. Suppose we redesignate L in figure 1 to be the supply of people capable of running Chinese restaurants and we lump all other factors of production under the heading of K. In that case, the immigration of ΔL causes a reduction Δw in the incomes of proprietors of Chinese restaurants and a corresponding gain of $L\ \Delta w$ to the recipients of the marginal products of all other factors of production.[4]

Another argument which can be placed in the same mold is that foreign labor is complementary and therefore beneficial to local labor because foreign workers do unpleasant or poorly paid jobs that the locals are not prepared to do. I find this argument difficult to accept for two reasons. The first is that the premise is not strictly true. Some of the original residents will continue to perform the unpleasant and poorly paid tasks in society; and those people, who would normally be among the poorest in the community, are the ones whose wages are held down by competition from migrants. Second, the complementarities between migrants and original residents will disappear in time. The first generation of migrants may be a special kind of labor but their children represent an ordinary addition to the labor force. The complementarity can only be preserved by the steady introduction of new migrants who will in time lose their distinctive characteristics. The short-term advantage is procured at the cost of a gradual expansion of the labor force.

Foreign ownership of capital.—The original residents of the migrants' country of destination may be adversely affected if part of that country's capital stock is foreign owned. Suppose once again that migration is pure labor ($\Delta K = 0$, ΔL is finite), but now add the assumptions that the total stock of capital, K, consists of K_1 owned by residents of the country to which the migrants come and K_2 owned by foreigners ($K = K_1 + K_2$) and that migration will not provoke new foreign investment. In this case, migration might lead to a reduction in the income of the original residents. The change in their income is

---

[4] Note that the area under the marginal product of labor curve in fig. 1 is actually the difference between the national income when an amount L is provided and the national income as it would be if none of the factor L were provided.

$$L\,\Delta w + K_1\,\Delta r = (-\tfrac{1}{2}\,\Delta L\,\Delta w) + (K_2\,\Delta r), \tag{7}$$

where the first term in parentheses is positive, the second term (a transfer to foreign owners of capital) is negative, and one cannot say a priori which is the larger of the two. This argument is taken directly from the international trade literature (MacDougall 1960).

The alleviation of shortage.—Our final variant of the terms of trade argument is that migration might be the means by which a country alleviates a shortage of a particular kind of labor. Suppose there is a substantial outward shift in the demand curve for, say, pipe fitters so that the actual wage becomes substantially in excess of the long-run "equilibrium" wage, where equilibrium is understood to mean that the wage is just sufficient to induce a potential pipe fitter to undertake the schooling and training necessary to become one. Here L in figure 1 is the number of pipe fitters prior to migration and ΔL is the number of pipe fitters migrating into the country. The point is that migration which is typically a small proportion of the total labor force in a given year (so that ΔL is small relative to L and the triangle ABC is small as well) may add significantly to a particular trade in short supply so that the relevant triangle ABC may be substantial. My colleague, Alan Green, after an intensive study of Canadian immigration policy, has come to the conclusion that this consideration goes a long way toward explaining the evolution of migration policy in Canada (see Green 1976).

The Distribution Argument

It is easily seen from figure 1 that migration of labor causes a redistribution of income from labor to capital. The size of the redistribution is $L\,\Delta w$, which is exactly equal to $K\,\Delta r$ if the triangle ABC is small enough to be ignored. The redistribution of income played no part in the analysis of the standard case and the terms of trade argument because it was taken as a premise that migration is to be counted as good or bad according as the total income of the original residents is increased and decreased. However, society may not be indifferent to distribution. If suppliers of labor are on the whole relatively poor and owners of capital are on the whole relatively rich, the migration of labor may be counted as harmful insofar as it causes a redistribution from labor to capital.

The Factor Market Argument

The general form of this argument is that migration is beneficial or harmful to original residents because the remuneration of the migrant's factors of production is less or more than the sum of their marginal products.

Progressive taxation.—Define the "net wage" to be the actual wage less tax payments plus the worker's share of public services financed through the tax system. With a genuinely progressive tax system, where the progressivity is not eroded through loopholes, the net wage rate defined in this special sense is greater than the marginal product of labor for the poor but less than the marginal product of labor for the rich whenever the gross money wage is equal to the marginal product of labor as assessed by the employer. Clearly, it is the net wage, rather than the actual money wage, that is relevant for the assessment of the costs and benefits of migration. It follows that progressive taxation tends to render migration beneficial or harmful to the original residents according to whether the migrants are relatively rich or poor.

Economies of scale.—With increasing returns to scale in the production function, the output Q is less than what is required for each factor of production to receive the value of its marginal product; increasing returns to scale in the production function $Q = F(K, L)$ implies that

$$Q < L\frac{\partial Q}{\partial L} + K\frac{\partial Q}{\partial K}, \tag{8}$$

so that the wage of labor is less than the marginal product of labor, or the rent of capital is less than the marginal product of capital, or both. If increasing returns to scale cause the wage to fall short of the marginal product of labor, then migration of an extra worker confers a benefit upon the rest of the economy. Of course the converse is also true. If decreasing returns to scale cause the wage to exceed the marginal product of labor, then migration of an extra worker results in a net loss to the rest of the economy.

Defense, looked upon as a pure public good, is a source of economies to scale to which the foregoing argument applies exactly. Migrants contribute to the cost of defense, reducing the cost to each of the original residents without reducing the amount of defense provided. On the other hand, land, minerals, water, and space, which are all in fixed supply, may be sources of decreasing returns to scale and a potential cost of migration to the original residents. One cannot say a priori where the balance lies.

Externalities.—Migration is advantageous or disadvantageous to the original residents of the country where the migrant goes if the characteristics of the migrant are such that his activities confer externalities or disexternalities on the people around him.

There are several variants of this argument, the most important having to do with the migration of highly educated people. It is often alleged—though I have seen no convincing evidence for this proposition—that the net contribution of educated people is in excess of their wage. If so, then inmigration of educated people is beneficial and outmigration of educated people is harmful to the rest of the population.

Migrants would also convey externalities if their propensity to save were greater than the propensity to save of the original residents and if—as is almost always the case—the tax laws were such that the social return to saving is significantly in excess of the private rate of return (Rodriguez 1975). It has been alleged that the inflow of migrants can affect excess demand or the balance of payments (Mishan and Needleman 1966, 1968). Similarly and more directly, migrants could be said to convey a kind of externality upon the rest of the population if their crime rates were lower or if they were inclined toward professions for which social benefits are in excess of private remuneration.

The cost of education.—The cost of education would be irrelevant to the assignment of gains and losses from migration if each man paid the full cost of his education, but it becomes important when education is subsidized or provided free of charge by the state. It is sometimes supposed that there is an implicit contract between the student and the state in which the latter supplies education at less than cost on the understanding that the net income of educated labor will, one way or another, be less than its marginal product. The immigration of educated labor generates the benefits of this arrangement without the cost. This argument is really a variant of the externalities argument above. Its validity rests entirely on the assumption that educated labor is paid less than the value of its marginal product.

The converse of this argument—that the exit of educated labor is harmful—has received a great deal of attention and is probably stronger, especially in the context of the migration of educated labor from less developed to developed countries. If, for whatever reason, the government wishes to maintain a given proportion of educated people, then the outmigration of educated people forces the government to provide more education than would otherwise be necessary. Two people must, as it were, be educated to ensure that one educated person remains in the country, and the education of the extra person, the one who migrates, is a social cost with no corresponding social gain unless, as Johnson (1965, pp. 304–6) has argued, the migrants make large transfers to their countries of origin or add significantly to the stock of knowledge in a way that is as likely to benefit one region as another. The cost of education plays an important role in the analysis of the brain drain in the writings of Johnson, Bhagwati, and others cited above.

Public Property

The argument we are about to develop can be classified as a special case of the factor market argument.[5] It is a characteristic of most societies today that a large portion of the property is publicly owned. This includes not only property that is officially in the public sector—public housing, roads, schools, national parks, public industry, and the like—but also the public's share of property that is officially private—the half of all capital in the corporate sector that accrues to the government via the corporation income tax, the present value of the revenue from inheritance tax, and so on. Even human capital is public insofar as the earnings from human capital are assessed at a very high rate of tax. On going from one country to another, the migrant abandons his share of the public property in the country from which he comes and acquires his share of public property in the country to which he goes.

Let us reconsider the case discussed above where migrants contribute ΔL to the labor force ($\Delta K = 0$), but let us suppose that a proportion θ of the capital stock is publicly owned and that the migrants acquire a share of public capital in their adopted country. The migrants' remuneration from publicly owned capital is $\theta(r + \Delta r)K[\Delta L/(L + \Delta L)]$, which is the product of the public's share of the capital stock θ, the income from capital $(r + \Delta r)K$ and the percentage increase in the labor force $\Delta L/(L + \Delta L)$. Equation (2) may now be revised as follows:

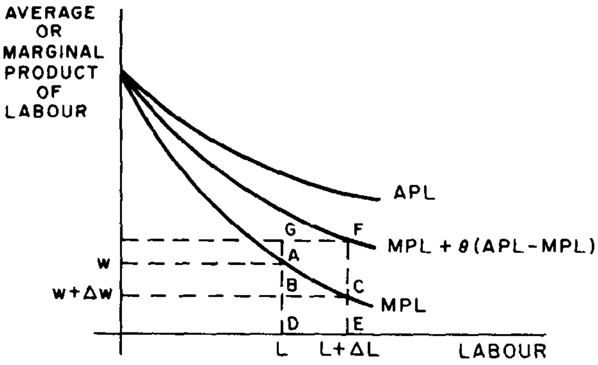
$$Q + \Delta Q = (wL + rK) + \{L\,\Delta w + K\,\Delta r - \theta(r + \Delta r)K[\Delta L/(L + \Delta L)]\} + \{(w + \Delta w)\,\Delta L + \theta(r + \Delta r)K[\Delta L/(L + \Delta L)]\}, \qquad (9)$$

where, once again, the three expressions on the right-hand side of the equation represent the income of the original residents prior to migration, the change in their income as a result of the migration, and the income of the migrants themselves. As equation (3) still holds, the value of $L\,\Delta w + K\,\Delta r$ is equal to $(-\tfrac{1}{2}\,\Delta w\,\Delta L)$ as before, and the net change in the incomes of the original residents is $[\tfrac{1}{2}\,\Delta w - \theta K(r + \Delta r)/(L + \Delta L)]\,\Delta L$, which could be positive or negative but would normally be negative if θ were at all large. The meaning of this expression can be illustrated on a diagram similar to figure 1.

Figure 2 reproduces the marginal product of labor (MPL) curve from figure 1 and adds two other curves, the average product of labor curve (APL) and a cross between average and marginal product represented

---

[5] In fact, Romans (1974) has gone so far as to claim that every argument about the cost and benefits of migration is a factor market argument. The terms of trade argument can be forced into that mold, though I do not think it useful to do so. Only the distribution argument is genuinely different.



ABC = 1/2 ΔW ΔL = gain to original residents if $\theta = 0$

BCFG = $\frac{\theta K (r + \Delta r) \Delta L}{L + \Delta L}$ = loss to original residents due to transfer to migrants of a share of publicly-owned capital.

GFCA = BCFG - ABC = net loss to original residents. (Notice that GFCA could be negative if θ were very small.)

ADEC = marginal product of migrants.

GDEF = full income of migrants.

Fig. 2

by the formula $MPL + \theta(APL - MPL)$. The third curve shows the income per migrant inclusive of his share of publicly owned capital, for

$$(w + \Delta w) + \frac{\theta K(r + \Delta r)}{L + \Delta L} = MPL + \theta(APL - MPL). \qquad (10)$$

One can immediately see from the diagram that the income of migrants exceeds their net contribution to national product if θ is large, for their contribution to national product is *ADEC* as before while their income is *GDEF*, equal to $\Delta L[MPL + \theta(APL - MPL)]$, and the difference *GACF* has to show up as a reduction in the remuneration of original residents.

Though the analysis is presented as comparative statics, the gains and losses can be capitalized and expressed as present values. There is in fact a good reason for doing so. If I migrate to a country, I impose costs upon the original residents of that country this year and in every future year. It seems natural to represent those costs as the sum of money that would just compensate the original residents. In all this, there is an implied demography. It is assumed that migrants come in families which replicate the age and sex distribution of the original population, and that the rates of birth, death, reproduction, and even outmigration of original residents and migrants will henceforth be the same. It is because migrants are exactly like the original residents that we can limit the impact of migration to its effect on the capital-labor ratio and on the pattern of ownership of property.

Some Back-of-the-Envelope Calculations

It is possible to attach rough magnitudes to this analysis. I shall attempt to compute the values of the different costs and benefits to the original residents of the migration of 1 million people into the United Kingdom in the year 1974. It will first be assumed that migrants have no capital of their own, and afterwards a calculation will be made of how much capital the migrant would need to bring with him in order that costs and benefits to the original residents would just cancel out.

All the data we require are presented in table 1. As the reader can see at a glance, some of the data are probably quite accurate, and some are most certainly not.

The classification of property requires a word of explanation. There are three main categories, public, private, and mixed. Public property is that which is owned outright by the state, including roads, bridges, hospitals, schools, etc., and an arbitrary imputation, entitled public resources, for items like crown lands, subsoil assets, and oil reserves. Mixed property is capital that is privately owned and managed but taxed in such a way that a substantial proportion of the income accrues to the state. I have assumed the public's share of the revenue from mixed property to be 50 percent, and, because of the corporation income tax, I have counted the assets of corporations as mixed property ignoring the exemption for interest payments. Private property is that for which the entire earnings accrue to the owner.

In assessing the value of the transfer from original residents to migrants associated with the migrants' acquisition of a share of publicly owned assets, we must, of course, deduct the value of public debt because the migrants confer a benefit upon the original residents by assuming a share of the taxes required to finance the interest and redemption of the debt. This could in principle outweigh the advantages the migrant receives from the acquisition of a share of publicly owned assets, but, as it happens, the debt is quite small compared with the value of publicly owned assets, and the balance is strongly in the migrant's favor. The measure of the debt is an estimate of its market value, which is substantially less than book value because interest rates were higher in 1974 than they were when the debt was issued. Thus of total assets of £304 billion, the amount that is publicly owned can be assessed in billions of pounds as:

Public assets	163
Half of mixed assets	39
Less public debt	−24
Total	178

TABLE 1

DATA FOR THE UNITED KINGDOM, 1974

1. Total capital stock:	£304 billion	4. Population	56 million
a) Private assets	£ 63 billion	5. Labor force	26 million
b) Mixed assets, foreign	£ 10 billion	6. Elasticity of substitution between labor and capital	1.0
c) Mixed assets, domestic	£ 69 billion		
d) Public assets	£113 billion		
e) Public resources	£ 50 billion	7. Labor's share of output	3/4
2. Public debt	£ 24 billion		
3. National income	£ 66 billion	8. Assumed migration	1 million

SOURCE.—Main source of information is *National Income and Expenditure*, 1964–74 (London: HMSO, 1975). National income (3) is from table 1, capital stock figures are from table 72, but the total (1) is augmented by a figure of £50 billion, included in this table as public resources (*e*) and representing the public's claim to subsoil assets, public land, etc. Public assets (*d*) consists of the categories "public corporations," "central government," and "local government." Private assets (*a*) consists of "personal sector" and "financial companies." The category of mixed assets (*b*) and (*c*) consists of "industrial and commercial companies," of which £10 billion is assumed to be foreign owned; this last figure is taken from the *Bank of England Quarterly Bulletin* (June 1974). Population (4) and labor force (5) are taken from the *Annual Abstract of Statistics* (London: HMSO, 1975). Labor's share of output (7) is actually lower than the actual figure for 1974 but higher than has been the case in recent years. The estimate of public debt (2) is of its market value in 1974, which is somewhat lower than its book value in 1974

Publicly owned assets amount to £178 billion or 58 percent of total assets, so that the value of θ in equation (9) is 0.58. Since publicly owned property is by definition owned equally by all citizens, the inflow of a million migrants involves the original residents in a loss of $\frac{1}{57}$ of the public property (the population of the United Kingdom was 56 million) or £3,123 per migrant per year.

To connect this measure of loss per migrant through the sharing of public property with the discussion in the preceding section and to compare it with other sources of gain or loss from migration, it is convenient to recompute the number in a roundabout manner, expressing gains or losses as annual flows in the first instance and capitalizing these. The calculations are presented in table 2 which is constructed on the assumption that the U.K. economy can be represented by an aggregate production function with unitary elasticity of substitution between labor and capital.

Table 2 is largely self-explanatory. The first row shows amounts of factors of production, the national income, and its components prior to migration; the second presents this information as it would be after the inflow of a million immigrants; the third row shows the effect of migration on total income, the income of migrants, and the income of the original residents; and the fourth row capitalizes the values in the third row at the rate of interest implicit in the example.

In interpreting this table, the following points should be borne in mind: (*a*) The immigration of a million people increases the labor force by only 464 thousand because the labor-participation rate of migrants is assumed to be the same as that of the original population. (*b*) The total income accruing to labor is apportioned between immigrants and original residents according to their shares of the total population. (*c*) The total

TABLE 2

THE EFFECT ON THE NATIONAL INCOME AND ITS DISTRIBUTION OF THE IMMIGRATION OF ONE MILLION PEOPLE INTO THE UNITED KINGDOM

				LABOR INCOME (£ million)		RETURN TO PRIVATELY OWNED CAPITAL (£ million)		RETURN TO PUBLICLY OWNED CAPITAL (£ million)	
	LABOR FORCE (million)	CAPITAL STOCK (£ million)	NATIONAL INCOME (£ million	Accruing to Original Residents	Accruing to Migrants	Accruing to Original Residents	Accruing to Foreigners	Accruing to Original Residents	Accruing to Migrants
Before migration .	26	304,000	66,000	49,500	.	6,567	271	9,662	
After migration	26 464	304,000	66,881	49,281	880	6,655	275	9,618	172 $\theta K(r + \Delta r)$ $[\Delta L/(L + \Delta L)]$
Change resulting from migration .	464 ΔL		881 ΔQ	−219 $L\,\Delta w$	880 $(w + \Delta w)\,\Delta L$	88 $(1 - \theta)K\,\Delta r$	4	−44 $\theta K\,\Delta r$	172
Present value of change resulting from migration	.		16,018	−3,981	16,000	1,600	73	−800	3,123 $\theta K\,\Delta L/(L + \Delta L)$

NOTE.—The numbers in table 2 are either reproduced from table 1 or computed from data contained there Items 6 and 7 in table 1 together imply that the aggregate production function is Cobb-Douglas with an elasticity of output with respect to labor of 3/4. Some numbers have been changed minutely to compensate for founding errors.

income accruing to capital is apportioned between the public and private sectors according to their shares of the ownership of the capital stock. Prior to migration, the £304 billion worth of capital consists of £178 billion publicly owned, £121 billion privately owned by residents, and £5 billion foreign owned (£5 billion rather than the £10 billion shown in table 1, because foreign capital is classified as mixed capital). After migration, one fifty-seventh of the £178 billion of publicly owned capital is assumed to be owned by migrants. (*d*) The appropriate rate of interest for converting the annual costs and benefits in the third row to the present values in the fourth is 5.5 percent, equal to the ratio of the total return to capital and the value of the capital stock. Note that this procedure is valid only if we can assume that the migration has no effect on the relative price of capital goods and consumption goods. (*e*) The numbers in the bottom row show at once the effects of the inflow of a million immigrants in millions of pounds and the costs or benefits per immigrant in pounds. The figure in the extreme lower right-hand corner of the table indicates that each migrant's share of the present value of the return from public property is £3,123; this is the same number that has already been computed as a fifty-seventh part of the total of public property.

The overall effect of migration upon the original residents can be computed either (i) by comparing the migrants' remuneration from labor and from publicly owned capital with increases in the gross domestic product resulting from the addition of the migrants' factors of production or (ii) by adding up the changes in the returns to the factors of production owned by the original residents. In the first procedure (adding up the annual changes from the third row), the values expressed in millions of pounds are:

Migrants' remuneration:		
Return to migrants' labor	880	
Return to migrants' share of publicly owned property	172	
Total		1,052
Increase in the national income:		
Increase in national income	881	
Less increase in return to foreign-owned capital	−4	
Total		877
Loss to original residents		−175

In this calculation, the loss to the original residents is completely dominated by the migrants' share of publicly owned property. The difference between the increase in the national income and the labor income of migrants, the famous triangle *ABC* in figures 1 and 2, is only £1 million, and the increase in the return to foreign owners of capital is also small. In the second procedure we add up the changes in the remuneration of the original residents:

Labor income	−219
Private return to capital	88
Return to publicly owned capital	−44
Total loss to original residents	−175

It should be stressed that the appropriate costs and benefits of migration are the present values in the fourth row of table 2 rather than the annual flows in the third row, for migration is a once-and-for-all event with consequences that persist indefinitely into the future. The total present value of the cost of migration associated with the sharing of public capital, the change in the remuneration to foreign owners of capital, and the terms of trade effect is the present value of £175, or £3,182 per migrant.

Even within the terms of this example, the result that the cost to the original residents is £3,182 per migrant could be overturned if the migrant was a supplier of publicly owned capital. It is not sufficient that he brings capital with him that would only reduce the size of the welfare gain due to the spread between the marginal products of the first and last migrant. To count as publicly owned, his capital must be invested so that it becomes mixed capital as defined above; the migrant must invest his capital in a form where a substantial part of the return to the capital accrues to the state, as would be the case if migrants subscribed to new issues of capital stock. If half the migrant's capital is private and the other half is mixed, he would need a capital stock of £12,728 per migrant or £50,912 per family of four in order that the original residents would be neither worse off nor better off as a result of his coming. Similarly, the figure of £3,182 can be looked upon as the present value of the benefits that the migrant would need to confer upon the rest of the community—by paying taxes high enough to cover the imputed rent of publicly owned property as well as the direct cost of the public services he receives or by conferring externalities that flow automatically from his scientific or technical training without being reflected in the wage that he receives—to overbalance the effect of his acquisition of a share of publicly owned capital.

General Comments

1. We have discussed a number of ways in which migration can have a positive or negative effect upon the welfare of the people who remain behind in the migrant's country of origin and of the original occupants

of the country to which he goes. We have of course emphasized the public property argument because that is what this article is about. But the assessment of the effects of any particular migration, such as the postwar migration into Canada or the migration of educated people from less developed to developed countries, would ideally take account of all the "exceptions" we have discussed other than those which can be shown to be irrelevant or quantitatively insignificant.

In such an assessment, it would have to be recognized that the exceptions differ in their timing. The analysis as illustrated in figure 1 is comparative statics. It may be interpreted in the short run as pertaining to a situation which lasts for 1 year or for a limited period of time. It may be interpreted in the long run as pertaining to a situation which persists indefinitely. The public property argument is intrinsically long run. Migrants are assumed to behave like the original residents of their country of adoption with the result that the capital-labor ratio of the original residents is permanently lower unless the original residents undertake enough extra saving to restore the capital-labor ratio to what it would have been in the absence of migration. It is because the public property argument is long run that we capitalized the values of the relevant areas in figure 2 in our calculation above.

Other considerations are short or medium run. Presumably migrant-specific factors of production cease to be migrant specific after the lapse of a generation. Shortages solve themselves in time. The children of poor migrants might be rich, and the children of rich migrants might be poor. Of course, both long- and short-term costs and benefits can be capitalized, but the appropriate weightings are different in each case.

2. The assessment of the consequences of the migration of educated labor from less developed to developed countries—the analysis of the brain drain—has for the most part been based on the externalities argument and the cost of education argument, and the general conclusion seems to have been that the less developed countries lose and the developed countries gain from migration. That conclusion is not necessarily overthrown when we introduce the public property argument, but it needs to be qualified.

First, the costs to those who remain behind in the migrant's country of origin and the benefits to original residents of the country to which he goes—costs and benefits having to do with the public expenditure on his education or with a divergence between his wage and the marginal product of his labor—will materialize only if the migrant is better educated than the average person in both populations. For it follows as an accounting identity that richer and better-educated people cannot receive less than the value of the marginal products of their factors of production unless the poor and worse-educated people receive more. Furthermore, it is not sufficient that the migrant himself be well educated; any assessment of the externalities resulting from the migrant's education must take account of the education of the migrant's dependents. A migrant can be said to convey benefits upon the original residents only if the average education of the migrant and his dependents together exceeds the average in the population as a whole.

There is even a possibility that the educated migrant might impose costs both upon the original residents of the country to which he goes and upon the people who remain behind in the country from which he comes. This could occur if the first country is well endowed with publicly owned capital while the second is heavily burdened with public debt, or if the migrant's human capital (assuming that human capital is partly public) were less than the average endowment of human capital in the first country but greater than that in the second. Both these situations could arise as a result of migration from poor to rich countries.

Second, since the public capital argument applies to rich and poor, to educated and uneducated migrants alike, a migration will not be beneficial to the original residents unless the externalities associated with the migrants' education or wealth more than compensate for the loss associated with the sharing of public property. There is no question that the United States benefited from the migration of Albert Einstein. It is difficult to say how much less skilled one can be before the net benefit to original residents is lost

Third, a migrant may convey benefits to some of the remaining residents of his country of origin if he is unable to take all of his possessions with him when he migrates. He confers benefits upon his brothers and sisters if he must abandon his share of the family inheritance. The obvious case is where a farm is entailed to be inherited as a whole by someone who is prepared to occupy it. The family property argument is a variant of the public property argument. Its importance depends on how large a portion of a man's capital he can maintain as he moves from one country to another.

3. Given the assumptions we have made, the original residents of a country could avoid the potential cost of migration or turn migration to their advantage in several ways: by the imposition of an entrance fee, by discriminating among migrants so as to accept only those who are rich, highly educated, or highly skilled, by increasing the public debt to the point where publicly owned assets are just matched by public liabilities, or by admitting migrants as guest workers rather than as citizens. The first method would oblige migrants to buy out a share of publicly owned capital, but the "optimal" entrance fee from the point of view of the original residents might be higher than this objective would imply. The second method is widely practiced. For instance, Canada has a point system in which education and skill are weighted heavily. Other countries have established themselves as tax havens by restricting migration to

people whose incomes are in excess of some given amount. The third method would not work in practice because debt can be repudiated directly or through inflation, and because the temptation to repudiate all or part of the debt would increase with the size of the debt itself. On the other hand, countries have admitted guest workers and may well have increased the incomes of the original residents by doing so. That is not inconsistent with the analysis presented here because it is the essence of the institution of guest workers that they are denied access to at least a substantial part of the advantages available to full citizens—support in old age, education for their children, the right to vote, and so on. This essay is an analysis of migration on the assumption that migrants become full citizens and that they and their descendants are as likely to remain in the country as the original residents themselves.

4. There is a close family resemblance between the economics of migration and the economics of birth control, for both have to do with the impact upon the original residents of a country of the introduction of extra people with an immediate claim to a share in society's capital stock. In fact, the economics of birth control is really the simpler of the two because it does not require that one identify the extent of public ownership in the capital stock. The newly born child acquires a share in the entire capital stock at the expense of his brothers and sisters if the capital stock is privately owned.

Conclusion

The main proposition in this paper can be put very simply: We can think of people in society as economic men and as political men. Economic men are unequal, for their incomes depend on the marginal products of the factors of production at their disposal and factors of production are not divided up equally. Political men are (or ought to be) equal, for each man has one and only one vote, the same civil rights as every other man, the same influence in the choice of leaders, and the same right to stand for office. But the equality of political men and the inequality of economic men cannot be kept in watertight compartments. Political equality exerts a continual pressure toward economic equality through public provision of services, progressivity in the tax system, and partial public ownership of the means of production. Migration is not just the transfer of factors of production from one country to another. The migrant acquires citizenship which entitles him to a fair share of public property and of the benefit or cost of redistribution, depending on whether he is rich or poor. This is an important consideration in the assessment of the effect of migration on people who remain behind in the country from which the migrant comes and on the original residents of the country to which he goes.

References

Berry, R., and Soligo, R. "Some Welfare Aspects of International Migration." *J.P.E.* 77, no. 5 (September/October 1969): 778–94.

Bhagwati, J., ed. *The Brain Drain and Taxation. II. Theory and Empirical Analysis.* Amsterdam: North-Holland, 1976.

Bhagwati, J., and Partington, M., eds. *Taxing the Brain Drain. I. A Proposal.* Amsterdam: North-Holland, 1976.

Bhagwati, J., et al. "The Brain Drain: A Symposium." *J. Development Econ.* 2 (September 1975): 193–318.

Borts, G. H., and Stein, J. L. *Economic Growth in a Free Market.* New York: Columbia Univ. Press, 1964.

Ferguson, C. E. *The Neoclassical Theory of Production and Distribution.* London: Cambridge Univ. Press, 1969.

Green, A. *Immigration and the Postwar Canadian Economy.* Toronto: Macmillan, 1976.

Grubel, H., and Scott, A. "The International Flow of Human Capital." *A.E.R. Papers and Proceedings* 56 (May 1966): 268–74.

Hamada, K. "Taxing the Brain Drain: A Global Point of View." In *The New International Order,* edited by J. Bhagwati. Cambridge, Mass.: M.I.T. Press, 1977.

Johnson, H. G. "The Economics of the Brain Drain: The Canadian Case." *Minerva* 3 (Spring 1965): 299–311.

———. "Some Economic Aspects of the Brain Drain." *Pakistan Development Rev.* 7 (August 1967): 379–411.

MacDougall, G. D. A. "The Benefits and Costs of Private Investment from Abroad: A Theoretical Approach." *Econ. Rec.* 36 (March 1960): 13–35.

Mishan, E. J., and Needleman, L. "Immigration, Excess Aggregate Demand and the Balance of Payments." *Economica,* n.s. 33 (May 1966): 129–47.

———. "Immigration: Some Long Term Consequences." *Econ. Internazionale* 21, pt. A (March 1968): 281–300, and ibid. pt. B (August 1968): 515–24.

Rodriguez, C. A. "On the Welfare Aspects of International Migration." *J.P.E.* 83, no. 5 (October 1975): 1065–72.

Romans, T. "Benefits and Burdens of Migration (with Special Reference to the Brain Drain)." *Southern Econ. J.* 40 (January 1974): 447–55.

The Monetary Appreciation of Paintings

John Picard Stein

Rand Corporation

Price indices for pre-World War II paintings are developed from U.S. and U.K. auction prices, assuming that auctioned paintings are sampled from a fixed underlying stock of paintings. The simple form of the capital-asset pricing model is modified to describe durable goods with risky capital appreciation and constant service yield. Estimated parameters for U.S. paintings are $\beta = 0.82$ (SE $= 0.34$), $R^2 = .24$, service yield $= 1.6$ percent (SE $= 7$ percent). To individual collectors, "performance" is positive to the extent that they value the service yield above 1.6 percent. Paintings are "inefficient." Regressions using U.K. data have no significant β coefficients.

In a number of ways, paintings are extraordinary economic goods. They are at once durable consumer goods and financial assets. Their aggregate supply is nonaugmentable, when the works of only deceased artists are considered. They are extremely heterogeneous, with values that range from a few dollars to a few million dollars. When hung in museums, they are public goods in the sense that their viewing services are simultaneously available to everyone. They are speculative goods to the extent that demand determines future price appreciation and expected future price appreciation determines demand. In sum, paintings are an archetype of what might be called "collector goods," and as such they deserve attention.

This study attempts to quantify and then to analyze the monetary appreciation of paintings over time. It is widely believed that the rate of return from investments in paintings during the 1950s and 1960s was far greater than the rate obtainable from comparable risky assets in financial markets (see Rush 1961, p. 385; Keen 1971, p. 24). The principal obstacle to informed discussion of this issue is the lack of a reliable measure

[*Journal of Political Economy*, 1977, vol. 85, no. 5]

© 1977 by The University of Chicago. All rights reserved

of the general rise in the price of paintings. In the section that follows, I develop price indices by assuming that paintings at auction are randomly and independently sampled from an underlying fixed stock of auctionable paintings. Then the auction prices of individual paintings constitute observations from the underlying distribution of painting values, and price indices for the total stock can be based on observations of the prices of different paintings each year. I construct annual price indices for the United States and Britain for the years 1946–68. The fixed stock of paintings chosen for study includes all paintings by artists who died before 1946.

In the subsequent section, I extend the simple form of the capital-asset pricing model, developed in articles by Sharpe (1963, 1964) and Lintner (1965*a*, 1965*b*), to describe risky durable goods. The riskiness of paintings is argued to arise from fluctuations in their rate of price appreciation, while the rate of nonpecuniary return is assumed to be constant over time. This model is fitted to the price indices for paintings to derive estimates of the nonpecuniary return and the performance and efficiency of paintings as *financial* assets.

Price Indices for Paintings

To permit comparisons over time, a fixed underlying stock of auctionable paintings must be isolated, from which paintings at auction are sampled. If the underlying stock is allowed to grow or change, price indices based on successive auction samples require a quality adjustment. Of the groups of paintings whose supply was essentially fixed during the period under study, 1946–68, the largest is the stock of all pictures painted before 1946. In the following analysis, I took as my object of study the latter group, further restricted here to paintings by artists who died before 1946, to avoid situations in which an artist whose paintings were under study produced new and closely substitutable art works after 1946.[1] Henceforth the term "paintings" will refer only to this restricted set of art works.

Suppose that the stock of paintings consists of n paintings. If Dp_i/p_i is the average instantaneous rate of monetary appreciation of painting i over the arbitrary period Δt between points s and t, then the mean proportional appreciation of all paintings is

$$P_t/P_s = \exp\left[\frac{1}{n}\sum_{i=1}^{n}\left(\frac{Dp_i}{p_i}\right)\Delta t\right]. \qquad (1)$$

---

[1] The roles of substitutable art works as well as all other complements and substitutes for the paintings thus isolated for study are summarized in the capital-asset pricing model by the behavior of the market portfolio. Interdependence among asset returns unrelated to market risk is ignored.

If one substitutes $D(\ln p_i)$ for Dp_i/p_i and uses the limit definition of a derivative, it follows that

$$\frac{P_t}{P_s} \cong \left(\prod_{i=1}^{n} p_{it}\right)^{1/n} \Bigg/ \left(\prod_{i=1}^{n} p_{is}\right)^{1/n}. \qquad (2)$$

The ratio of the geometric means of prices in any two periods will be used as a price index.

Using the geometric mean of auction prices as a proxy for the geometric mean value among all paintings introduces two sources of bias. First, the subset of auctionable paintings is not fixed but decreases over time as top-quality works are skimmed off the market and accumulated in museums. Second, auction prices do not represent a uniform sample of the values of auctionable paintings. Some types of paintings are auctioned more or less frequently than others, and some auction prices may be biased by noncompetitive auction practices. A number of arguments are given below as to why these may not actually turn out to be problems; but in any case the data on auction prices are probably as good as one could hope to get.

Let us consider the bias introduced by museum collecting. To the extent that museum acquisitions draw proportionately from all quality segments of the auctionable stock, this depletion would neither skew the sample of auction prices nor bias the price indices as constructed above.

Post-1950 museum acquisitions could not begin to pinch the supply side of the auction market until near the close of the period of analysis. If one assumes that, on the average, privately owned paintings were turned over once in a generation during the 1950s and 1960s, then most of the paintings bought by museums after 1950 would have been unlikely to reappear on the market during the period of analysis, even if purchased by private persons. In contrast, museum growth prior to and during World War II did not take place on nearly so large a scale as later.[2]

During major recessions, an above-average number of persons liquidate portions of their assets and eventually come to sell paintings. During the sustained economic boom of the early and middle 1960s, probably fewer than the normal number of private collectors chose to liquidate their collections. Thus museums acted no differently from private individuals during those years—both held on to their collections.

Most of the very top-quality works suitable for museum collection are sold privately through negotiation rather than at auction. The museum accumulation of such works would not be missed by the auction market.

---

[2] Approximately half of the 2,372 U.S. museums operating in 1970 (art museums as well as all other types) were founded between 1950 and 1969, according to the American Museum Association, Washington, D.C. In addition to the accelerating rate of foundation of new museums, there is reason to believe that individual museums accelerated their pace of acquisition of pictures in the 1950s and 1960s.

Consider the second source of bias—the likelihood that auction prices do not represent a uniform sample of the values of auctionable paintings. Works of fashionable artists or schools may dominate the auction rooms out of proportion to their share in the total stock. Very expensive paintings are relatively unlikely to be auctioned, because of the sliding scale of commission fees charged by auction houses. The paintings of highest quality, as noted, are seldom sold at auction. Collusion among dealers (in what are known as "dealer rings") tends to depress auction prices. "Buy-backs," so called because a seller puts up a painting for auction and then purchases it himself through an intermediary in order to record an elevated value for the work which can later be a useful sales point, artificially inflate prices.

If it can be assumed that the net proportional bias caused by all these factors is randomly distributed about a mean that is constant over time, the net effect is to displace the geometric mean auction price from the average value of all auctionable paintings by the same expected proportion year after year. The relative appreciation of this mean will still be an unbiased estimate of the relative price index for all auctionable paintings.

The source of the auction data used here is *Art Prices Current*. The 1967-68 volume was the latest available for the study, and data-preparation and processing costs limited the analysis to the years 1946-47 to 1967 68. A screening of the transactions eliminated paintings in cases in which (1) experts in the auction houses had expressed doubt as to the painting's authenticity, (2) the painter was unknown or the painting was listed under "school of," or (3) the painter was still living in 1946.

Table 1 presents geometric-mean-price levels for sales in the United States and Britain, with sample sizes corresponding to each average. The U.S. and British art markets were studied because an average of 84 percent of the paintings reported in the data source were sold in these countries,[3] with the remainder being sold on the European continent. British auction prices were converted to U.S. dollars, using the exchange rate as of December 31, roughly midway through the art-market year, which runs from August to July. The 1947–49 volume of *Art Prices Current* combined the sales of the 2 years.

The average annually compounded rate of price appreciation for U.S. paintings between 1946–47 and 1967 68 was 10.47 percent. For dollar-denominated prices of U.K. paintings, this rate was 10.38 percent. For pound-denominated prices of U.K. paintings, it was 13.12 percent.[4]

---

[3] Each year's percentages weighted equally.

[4] I did not construct separate indices for various classes of paintings. Classification of individual paintings would have required an expert knowledge of art history, and in any case differences in the rates of appreciation among various classes of paintings are

TABLE 1

GEOMETRIC MEAN AUCTION PRICES

YEAR	UNITED STATES			BRITAIN			
	Geometric Mean Auction Price ($)	SE of Estimate (%)	*N* Paintings in Sample	Geometric Mean Auction Price (£)	Geometric Mean Auction Price ($)*	SE of Estimate (%)	*N* Paintings in Sample
1967–68	3,198	7.1	350	1,214	2,923	2.5	2,115
1966–67	2,300	6.7	258	829	2,314	2.5	2,176
1965–66	4,904	9.8	255	888	2,489	2.5	2,040
1964–65	3,087	10.2	190	1,011	2,821	2.9	1,887
1963–64	2,802	7.8	268	811	2,267	2.9	1,965
1962–63	2,445	8.3	321	810	2,270	3.4	1,580
1961–62	2,417	10.5	293	512	1,438	3.1	1,743
1960–61	1,185	8.1	374	355	995	2.8	2,240
1959–60	853	6.5	407	298	835	2.8	2,087
1958–59	1,057	9.3	370	234	656	3.1	1,998
1957–58	803	7.3	419	188	527	2.8	1,765
1956–57	623	5.8	442	190	529	3.3	1,528
1955–56	629	4.5	634	174	488	3.0	1,449
1954–55	583	5.0	475	142	396	2.6	1,097
1953–54	448	4.8	352	109	307	2.6	1,289
1952–53	536	5.8	402	133	373	3.4	1,145
1951–52	515	4.3	523	96.7	269	2.9	1,175
1950–51	529	4.3	501	107	299	2.8	1,470
1949–50	375	4.4	546	98.6	276	3.0	1,252
1947–49	454	3.6	837	98.1	395	2.2	2,297
1946–47	395	3.7	733	91.1	367	2.6	1,525

* Using exchange rates at December 31 of each year, as reported in *International Financial Statistics*

Random and systematic sampling errors cause the geometric mean auction price to differ from the geometric mean value within the underlying stock of paintings. The error variance of the sample average about the true average in any given year is $1/n$ times the price variance in the underlying population, where n is the number of paintings in the sample auctioned that year. Observed variance among auction prices in each year was used as an estimate of the price variance in the underlying population. The standard error of estimate of each geometric mean is given in table 1. It is expressed in percentage form as a result of working with logarithms of prices.

To check the stability over time of the process by which paintings were selected for auction, I graphed the frequency distributions of auction prices for the United States and Britain in each year. These distributions are typified in figure 1, rising abruptly from some "threshold" price (perhaps that price below which it was not profitable to auction a painting), reaching a maximum, and then decreasing with higher prices, consistent with the intuitive notion that paintings of increasingly high quality and value are increasingly scarce. The decile points of the frequency distribution as well as the geometric mean price and the threshold price were found by visual inspection to parallel each other closely over time.[5] A Chow test failed to show that the distribution of auction prices

---

largely due to fashion. *The Times* collaborated with Sotheby's auction house to develop price indices for selected classes of paintings and other art works in Britain over the period 1950-69. The following average annual rates of appreciation were observed for some of these classes: old master paintings, overall, 11 percent; northern primitives, 10 percent, Brueghel school, 11 percent; Dutch and Flemish flower pictures, 9 percent; Dutch and Flemish still lifes, 11 percent; Dutch marine pictures, 11 percent; Dutch landscapes, 11 percent, Dutch genre paintings, 9 percent; Dutch architectural paintings, 14 percent; Dutch peasant scenes, 8 percent; historical portraits, 8 percent; Italian primitives, 10 percent; early Renaissance paintings, 13 percent; late Renaissance paintings, 8 percent; seicento paintings, 15 percent; Italian eighteenth-century *vedutisti*, 13 percent; French eighteenth-century masters, 13 percent; French eighteenth-century view painters, 15 percent, pre-Raphaelite paintings, 15 percent; impressionist paintings, overall, 17 percent; Monet, 21 percent; Renoir, 13 percent; Pissarro, 17 percent; Sisley, 18 percent, Boudin, 18 percent, Fantin-Latour, 15 percent; twentieth-century paintings, overall, 21 percent; Picasso, 22 percent; Braque, 19 percent; Bonnard, 22 percent; Chagall, 24 percent; Vlaminck, 21 percent; Rouault, 16 percent; Utrillo, 17 percent; French furniture, 10 percent; English glass, 12 percent; English silver, 12 percent; old books, 14 percent; modern books, 11 percent; Chinese porcelain, 18 percent; old master drawings, 19 percent; old master prints, 20 percent. Observed prices were adjusted for the quality of the art work by Sotheby's appraisers before being incorporated in the indices. Many of the indices were based on a handful of observed prices. The construction of these indices is described in Keen (1971, pp. 242–49) and critiqued in Stein (1973, pp. 46–47). According to these indices, excellent returns could be earned by anticipating fashions. In this study, I try to avoid having to explain fashion swings by considering as broad a class as possible of paintings by artists whose reputations had generally matured and thus stabilized.

[5] A more thorough analysis of the frequency distributions of auction prices can be found in Stein (1973). Graphs of the frequency distributions observed in each country in each year are presented there in an appendix.

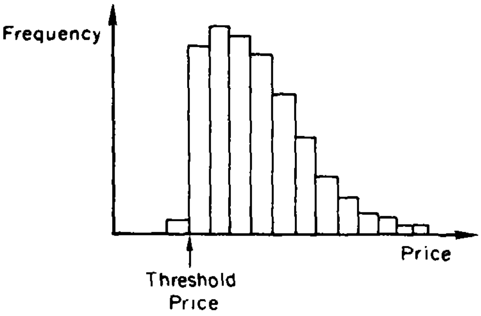


FIG. 1.—Typical frequency distribution of auction prices in a given country in a given year

around their mean was invariant over time for each country. In view of the large number of auction prices in each annual distribution, minor changes in either the sampling distribution or the underlying distribution of painting values could account for this negative result. Actually, for the price indices in this paper to be unbiased, it is only required that changes in the sampling distribution be nonsystematic over time.

Paintings as Financial Assets

Paintings are at once consumer durables, yielding a flow of nonpecuniary viewing services, and capital assets, yielding a return from financial appreciation. Thus, in a study of this kind, one may either modify a durable-goods model to incorporate positive capital appreciation and a special adjustment for risk or adjust a capital-asset model to account for nonpecuniary returns. The approaches are essentially identical, but the capital-asset model is more convenient because it explicitly accounts for risk.

The starting point used here was the simple form of the capital-asset pricing model developed by Sharpe and Lintner.[6] According to the ex post analogue of this model, known as the "diagonal model" or the "market model" (see Sharpe 1963; Fama 1968), under certain assumptions

$$R_{jt} - R_{Ft} = (R_{Mt} - R_{Ft})\beta_j + e_{jt}$$
$$\beta_j = \frac{\operatorname{cov}(R_j R_M)}{\sigma^2(R_M)}. \tag{3}$$

Here R_{jt} is the random, one-period rate of return on any asset j; R_{Mt} is the random, one-period rate of return on the market portfolio; R_{Ft} is the exogenously given, one-period rate of return on the risk-free asset; a parameter β_j, known as "relative systematic risk," characterizes each asset j; and e_{jt} is a random error term with a zero mean. Essentially the model in this form states that the observed return on any asset, expressed as a premium above the risk-free rate of return, is proportional, with a stochastic error term, to the asset's relative systematic risk, where the factor of proportionality is the observed premium of the market portfolio's return above the risk-free rate.

---

[6] See Jensen (1972) for a review of recent theoretical and empirical work on this model.

To evaluate the financial behavior of individual assets and portfolios, Jensen (1969) developed measures of "performance" and "efficiency." The performance of an asset is the difference between its observed return over a specified time period and the return that could have been expected ex ante on the basis of the asset's systematic risk. Adjustment is made for observed effects caused by the realized outcome of the unpredictable random factor in overall market behavior (known as the "market factor" in the market model). Assets are "superior," "neutral," or "inferior," according to whether average performance over time is greater than, equal to, or less than zero.

The efficiency of an asset is the difference between its observed return over a specified time period and the return that could have been expected ex ante on an alternative, naively constructed portfolio made up by combining the market portfolio and the riskless asset in such proportions as to achieve the same variance of return or "total risk" as in the original asset. Again, adjustment is made for the realized outcome of the market factor. Efficiency compares observed return with total risk, systematic and nonsystematic; whereas performance compares observed return with systematic risk only. Assets are "inefficient," "efficient," or "superefficient," according to whether average efficiency over time is less than, equal to, or greater than zero.

The Total Return to Paintings and the Regression Equation

The overall rate of return to paintings, r_p, unadjusted for risk or for inflation in the general price level, can be divided into two components: a return from nominal capital appreciation, R_p, and a residual return from durable services, R_s, where R_s summarizes the net return from viewing services, less insurance and maintenance costs, plus an annualized premium to account for any tax advantages associated with investments in paintings, less an annualized premium to account for the illiquidity of investments in paintings. The riskiness of holding paintings arises from the variability of R_p and R_s over time. However, as argued below, R_s can be assumed to be nearly constant, so that $\sigma^2(r_p) \cong \sigma^2(R_p)$ and $\operatorname{cov}(r_p, R_M) \cong \operatorname{cov}(R_p, R_M)$.

The annual cost of fire and theft insurance for art works is typically

0.5 percent,[7] but no direct estimates could be found for the remaining components of R_s. The annualized premiums to account for tax advantages and illiquidity costs are small, since these benefits and costs arise only when paintings change ownership, assumed here to be once each generation on the average. On an annualized basis, such costs should not amount to more than 1 percent, and since tax advantages and illiquidity costs appear with opposite signs, the net premium is probably close to zero. It seems reasonable to assume that each of these components of R_s is constant over time.

The return from viewing services can be conveniently thought of in terms of the aggregate stock of paintings. This return can be expressed as the ratio of the nominal value of the annual viewing services of this stock to the nominal value of the stock. The value of the viewing services can be assumed to depend on slowly changing variables: the amount and distribution of nominal wealth among collectors and any substitution between these viewing services and all other goods and services. Assuming that any change in the valuation of the fixed flow of real viewing services leads to a parallel change in the value of the stock of paintings, the ratio of these two values will be constant through time. Thus each component of R_s may be assumed to be constant over time. It is clear from table 1 that R_p can vary widely in percentage terms.

An upper-bound estimate of the return from viewing services is provided by rental rates in the art-rental market. The average annual rental charge is about 11 percent of appraised value, when ancillary charges for insurance, shipping, hanging, and periodic replacement of paintings are excluded.[8] The clients of art-rental firms are generally corporations which for tax reasons save by renting rather than buying paintings. Since individuals seldom rent paintings under these circumstances, 11 percent may thus be taken as an upper bound on the return from viewing services. Because the net value of the other components of R_s is close to zero, 11 percent can also be taken as an approximate upper bound on R_s.

Substituting $R_{pt} + R_s$ for R_{jt} in equation (3) and adding a constant term which incorporates R_s, we have

$$(R_{pt} - R_{Ft}) = \alpha + (R_{Mt} - R_{Ft})\beta_p + e_{pt}$$

$$\beta_p = \frac{\mathrm{cov}\,(R_p, R_M)}{\sigma^2(R_M)}. \tag{4}$$

---

[7] According to Wellemeyer (1972), only a few firms carry special fine-arts policies; their rates range from 0.17 percent to 1.0 percent of appraised value per year, with the average running around 0.5 percent.

[8] The following rental programs were considered: Art Council, Boston, Art Programs, Los Angeles; Boston Museum of Art; Liberty Leasing, Chicago; Whitney Museum, New York.

When equation (4) is used as the regression equation, the estimated constant term is an unbiased and efficient estimator of the combined effects of $-R_s$ and any superior performance of the asset, assuming that the capital-asset pricing model is applicable. The estimated slope coefficient is an unbiased and efficient estimator of the relative systematic risk of paintings.

Empirical Results

A number of alternative data sets were available for this study.

Set R_p. Three series of annual rates of appreciation for paintings were computed from the U.S. and the two British (dollar- and pound-denominated) price series presented in table 1. Annual continuously compounded rates of return were defined by $R_{pt} \equiv \log_e (P_{t+1}/P_t)$, where P_t is a series from table 1.[9] The R_{pt} describes a 1-year period from January 31 to January 31.

Set R_F.—Government bills, medium-term government bonds, and long-term government bonds in the United States and Britain were tested as risk-free assets, since data describing the yield on government bonds with 1 year to maturity were unavailable. The annual yields were obtained from *International Financial Statistics* on each asset as of January 31, then converted into annual continuously compounded rates.

Set R_M.—Three all-stock indices were tested in the position of a general market asset.[10] Only one of these, the Fisher investment-performance index, accounts for dividend returns. The Fisher index includes all stocks on the New York Stock Exchange, weighting each according to its stock value outstanding.[11] Also tested were the Standard and Poor's *Index of Corporate Share Prices* in the United States and the *Financial Times Ordinary Share Index* of stock prices in Britain, as reported in *International Financial Statistics*. These indices, which do not cover dividends, are based on a number of key securities in each country and weight each security equally. In all cases, annual continuously compounded rates of return were calculated for the period January 31 to January 31.

Five regressions were run using medium-term government bonds as the riskless asset. Prices of U.S. paintings were regressed on both U.S. market indices; prices of U.K. paintings were regressed on the U.K. market index. And because of the possibility that the London art market might be governed by U.S. buyers, U.K. art prices were regressed on each of the U.S. market indices.

---

[9] Using continuously compounded rates avoids the arbitrary choice of a period length, necessary when using periodically compounded rates. For changes in the price level of less than ±15 percent, the change in the log price is very close to the percentage price change

[10] The conditions under which all stock portfolio can be used in place of the market portfolio are discussed in Jensen (1972).

[11] For years prior to 1960, these indices are presented and discussed in Fisher (1966). Updated indices were obtained from the author.

Only one regression, that relating the price indices for paintings sold in the United States and the Fisher index for the New York Stock Exchange, had a β coefficient significant at greater than the 95 percent level. In that regression,

$$
\begin{array}{llll}
(R_p - R_F) = & -0.016 & +\ 0.822(R_M - R_F); & R^2 = .242 \\
\text{SE:} & (0.072) & (0.343) & \text{Durbin-Watson statistic} = 2.18 \\
t\text{-values (18 df):} & -0.22 & 2.40 & \\
\text{Variable:} & R_p - R_F & R_M - R_F & \\
\text{Mean:} & 0.074 & 0.110 & \\
\text{Variance:} & 0.095 & 0.034 &
\end{array}
\tag{5}
$$

This regression was the only one involving painting and stock indices for the same country and a stock index incorporating dividend returns; therefore we are justified in concentrating on this single regression in the following conclusions. The first 10 conclusions below relate to this regression. The eleventh conclusion presents the other four regressions.

1. The β_p is significant at the 97 percent level. This finding supports the assumption that there is a relationship between the return on paintings and that on the market portfolio.

2. If the simple formulation of the capital-asset model can be assumed to apply, then the estimated relative systematic risk of paintings, $\hat{\beta}_p = 0.82$, is less than unity, though not significantly so. Therefore paintings are probably a somewhat conservative investment in comparison with equities, with regard to systematic risk. Any simple comparison of average returns in the paintings and equities markets would probably handicap the paintings, because their lower systematic risk must be taken into account.

3. The correlation coefficient, $R^2 = .24$, indicates that the market portfolio accounts for only 24 percent of the variance of the return premium of paintings. Unbiased measurement errors in the dependent variable do not bias the estimates of α and β_p, but they increase the variance of the unexplained residuals and bias the estimated R^2 coefficient downward. Standard errors of the estimated geometric mean painting values from table 1 account for variance in $(R_p - R_F)$ of .017 on the average, leaving a variance of 0.078 in the dependent variable, of which the market portfolio explains 44 percent.

4. The estimated annual net return from durable services, $\hat{R}_s$, is $-\hat{\alpha} = 1.6$ percent. In comparison, the 10.5 percent average annual appreciation of paintings in the United States between 1946 and 1968 lends credibility to the assumption that collectors regard paintings as

capital assets and that the prices of paintings satisfy conditions in the capital market regardless of whatever conditions they also satisfy in the durable-goods market.

However, because the estimate of the constant term has a large standard error, the results are consistent at the 95 percent confidence level, with the net yield from durable services being valued anywhere from about -13 percent to $+15$ percent per annum. Specifically, the results are consistent with the viewing services being valued at 11 percent per annum, the upper-bound figure estimated from picture-rental rates; or at 0.0 percent, a lower-bound figure appropriate if collectors were interested exclusively in financial return.

5. The performance of paintings, $E(\delta_p)$, is measured by $\hat{\alpha} + R_s$. Using the upper-bound estimate of R_s leads to an upper-bound estimate of superior performance of about 9.4 percent per annum ($-0.016 + 0.11 = 0.094$). In contrast, to the speculative investor not at all interested in viewing pleasure, performance is inferior by about 1.6 percent ($-0.016 + 0.00 = -0.016$). In general, to any investor, estimated performance is superior (inferior) by the amount that the net value of durable services exceeds (falls short of) 1.6 percent per year.

6. The efficiency of paintings, $E(\gamma_p)$, is related to their performance as follows:[12]

$$E(\gamma_p) = E(\delta_p) - [E(R_M) - R_F]\left[\frac{\sigma(R_p)}{\sigma(R_M)} - \beta_p\right]. \tag{6}$$

Taking as an estimate of $[E(R_M) - R_F]$ the 1946-68 average annual premium of the return on the market portfolio above the medium-term government-bond rate and using the estimates of $\sigma(R_p)$, $\sigma(R_M)$, and β_p given in equation (5), one estimates the efficiency of paintings to be 9.9 percent less than their performance. Thus only if nonpecuniary viewing pleasure were valued above 11.5 percent per annum ($9.9 + 1.6 = 11.5$), as the rental market indicates it generally is not, would paintings be estimated to be an efficient investment.

7. The Durbin-Watson statistic of 2.18, compared with a d_u of 1.41 and d_l of 1.20 at the 5 percent level of significance, indicates nonsignificant negative autocorrelation of residuals. So the model is well specified with regard to this factor.

8. The regression results are insensitive to the choice of the risk-free asset. When the regression was rerun using treasury-bill rates and then long-term (10-year) government-bond rates, all estimated statistics were very close to those in equation (5). So the use of 3-year rather than 1-year bonds as the riskless asset introduces no important bias.

---

[12] The relationship between performance and efficiency is developed and discussed in Jensen (1969, pp. 200–202).

9. To test whether the behavior of the paintings market was stationary over time, the regression was rerun for the first and second halves of the sample period. To check for heteroscedasticity, the residual sums of squares of the two regressions were compared using an F-test and found not to differ significantly at the 5 percent level. The required F-ratio, $F(8, 8) = 3.35$, lies below 3.44, the upper 5 percent point of the $F(8, 8)$ distribution. Under the assumption of homoscedasticity, an F-test was used to test for equality between the two halves of the sample period for the intercept and slope parameters in the underlying relations. The required F-ratio is $F(16, 2) = 1.80$, which is well below 4.42, the upper 20 percent point of the $F(16, 2)$ distribution. If one assumes the joint hypothesis that the variance of the error terms and the intercept and slope parameters were the same for the two halves of the sample period, the probability of obtaining F-statistics as large as $F(8, 8) = 3.35$ and $F(16, 2) = 1.80$ is greater than $[1 - (1 - 0.05)(1 - 0.20)] = 0.24$. The two sample halves are not significantly different even at the 24 percent level. This finding is partly a result of the fact that neither the overall relation for the entire sample period nor the individual relations for the two sample halves fit very well.

10. To verify that the results do not depend on the arbitrary length of the time period used in the analysis, a regression was run using 2-year rates of return. The annual basis for the art-price indices, of course, prevented testing the model for periods shorter than a year. Because the number of observations is reduced to 10, the standard errors obtained are relatively large. The results were consistent with the assumption that the length of the period used for regression may be chosen arbitrarily, but they are not a strong confirmation of that assumption.

11. The other four regressions, relating the alternative series for painting prices and market returns, failed to show any significant relationships. The results are given in table 2. Replacing the Fisher index by the U.S. share-price index to explain U.S. painting prices decreased the confidence level at which the slope coefficient was significant from more than 97 percent to less than 90 percent and reduced the R^2 from .24 to .14. The estimated β coefficient increased to more than unity, but this can be explained at least partly by the neglect of dividend returns in the U.S. share-price index. The estimated constant term reversed sign but remained highly nonsignificant. In the regression relating British painting prices and British share prices, the slope coefficient was nonsignificant at the 70 percent confidence level. In two regressions relating British painting prices to U.S. stock-market indices, the slope coefficients were nonsignificant at the 40 percent confidence level.

Industrial share prices, if dividend returns are ignored, do not significantly explain auction prices in either the United States or the United Kingdom. This conclusion has particular reference to other studies of the

TABLE 2

REGRESSION PARAMETERS FOR NONSIGNIFICANT RELATIONSHIPS BETWEEN PAINTINGS AND MARKET ASSETS

	DEFINITION OF VARIABLES			REGRESSION PARAMETERS				
REGRESSION	Paintings (R_p)	Market Portfolio (R_M)	Riskless Asset R_F	$\hat{\alpha}$	SE($\hat{\alpha}$)	$\hat{\beta}$	SE($\hat{\beta}$)	R^2
I.	Index of $ prices of paintings auctioned in U.S.*	Standard and Poor's *Index* (U.S.)	Yield on 3-year U.S. government bonds	.013	.074	1.01	.594	.14
II. . . .	Index of £ prices of paintings auctioned in U.K.*	*Financial Times Index* (U.K.)	Yield on 5-year U.K. government bonds	.799	.041	.311	.305	.06
III.	Index of $ prices of paintings auctioned in U.K.*	Fisher index (U.S.)	Yield on 3-year U.S. government bonds	.072	.052	.089	.259	.01
IV. .	Index of $ prices of paintings auctioned in U.K.*	Standard and Poor's *Index* (U.S.)	Yield on 3-year U.S. government bonds	.072	.046	.143	.291	.01

* Computed from table 1

art market (Rush 1961; Keen 1971) which have informally compared painting prices with share prices. By ignoring dividend returns, such comparisons produce misleading conclusions, overestimating the return to paintings relative to the return to equities.

The weak relation between London auction prices and London share prices also reflects the fact that many non-British buyers patronize the London art market and the British stock market does not influence their demand. The weak relations between British auction prices and the indices for the U.S. stock market suggest that American buyers do not dominate the London art market. Many non-British buyers are also non-U.S. buyers.

Conclusion

According to estimates based on U.S. markets, paintings appear to be a fairly ordinary asset, considered simply as an asset. To the investor primarily interested in financial gain and valuing the return from viewing pleasure at only about 1.6 percent per year, paintings are no more or less attractive than other assets. They yield the going rate for their systematic risk. Contrary to a misbelief that lures some collectors, investment in paintings is not particularly lucrative. The expected appreciation from such an investment is only 73 percent of the return,[13] including capital gains, in the equities market. And contrary to another misbelief, the values of paintings are not secure during times of economic recession, being subject to the same ups and downs as the stock market. In addition, the return to paintings contains a substantial element of nonsystematic risk, a factor that should persuade collectors to keep most of their wealth in other forms. Any superior performance derivable from paintings can be attributed entirely to the viewing pleasure they provide, not capturable by speculators.

References

Fama, Eugene. "Risk, Return and Equilibrium: Some Clarifying Comments." *J. Finance* 23 (March 1968): 29–40.

Fisher, Lawrence. "Some New Stock-Market Indexes." *J. Bus.* 39, pt. 2 (January 1966): 191–225.

International Monetary Fund. *International Financial Statistics.* Vols. 1–20. Washington: IMF, 1948–70.

Jensen, Michael. "Risk, the Pricing of Capital Assets, and the Evaluation of Investment Portfolios." *J. Bus.* 42 (April 1969): 167–247.

———. "Capital Markets: Theory and Evidence." *Bell J.* 3 (Autumn 1972): 357–98.

Keen, Geraldine. *The Sale of Works of Art: A Study Based on the Times-Sotheby Index.* London: Nelson, 1971.

Lintner, John. "The Valuation of Risky Assets and the Selection of Risky Investments in Stock Portfolios and Capital Budgets." *Rev. Econ. and Statis.* 47 (February 1965): 13–37. (*a*)

———. "Security Prices, Risk, and Maximal Gains from Diversification." *J. Finance* 20 (December 1965): 587–615. (*b*)

Office of the Fine Art Trade Journal, ed. *Art Prices Current, 1946/47 to 1967/68.* London: Art Trade Press, 1947–68.

Rush, Richard. *Art as an Investment.* New York: Bonanza, 1961.

Sharpe, William. "A Simplified Model for Portfolio Analysis." *Management Sci.* 9 (January 1963): 277–93.

———. "Capital Asset Prices: A Theory of Market Equilibrium under Conditions of Risk." *J. Finance* 19 (September 1964): 425–42.

Stein, John Picard. "The Appreciation of Paintings." Ph.D. dissertation, Univ. Chicago, 1973.

Wellemeyer, M. "The Money in Old Furniture." *Fortune* (November 1972), pp. 65–66.

---

[13] Over the period 1946–68, the annually compounded nominal appreciation of paintings was 10.5 percent, while the annually compounded nominal return in the stock market was 14.3 percent.

Empirical Study of Scale Economies and Production Complementarity: The Case of Journal Publication

---

William J. Baumol

Princeton University and New York University

Yale M. Braunstein

New York University

An empirical study applying the theory of scale economies and interproduct complementarity. Several recently formulated concepts of the theory of scale economics are examined with the aid of estimates of cost functions, using data for publishers of scientific journals. In one case the observations are shown to fall near the outer limits of the range of declining costs, implying that neither amalgamation nor breakup of firms will reduce total cost. In another case the observations fall well within the subadditive range, implying that amalgamation can provide savings. The study confirms the applicability of the new theory of scale economies.

Recently, one of the authors of this paper has shown that in the multiproduct industries which populate the real world empirical study of phenomena related to scale economies runs into complications which seem not to have been recognized before. Specifically, if one wishes to show that it is cheaper for industry output to be produced by one firm than by two or more firms (the concept of subadditivity of the cost function), it is generally not enough to estimate the cost function in the neighborhood of current output levels. We will show here, however, that this and other

The work reported here was supported by the Division of Science Information of the National Science Foundation. We are most grateful to Kent Bjornstad for carrying out the calculations with dedication and skill. Very helpful comments were also made by Dietrich Fischer, M. I. Nadiri, and an anonymous referee. Last, but by no means least, we must thank Arthur Rosenzweig for saving us from making a rather serious error and Richard E. Quandt for help in choosing an appropriate test of significance for the parameter estimates in our nonlinear relationships.

[*Journal of Political Economy*, 1977, vol. 85, no. 5]

© 1977 by The University of Chicago. All rights reserved.

problems besetting empirical work in the area can be overcome. Using data on the cost of publication of scientific journals, we will show how these issues can be analyzed effectively.

Specifically, the problem is that, as is always true in reality (though not in the textbooks), we are dealing here with a multiproduct industry. Most of the usual scale-economies concepts do not apply to the multiproduct case without considerable amendment. Moreover, there seem to have been no previous statistical tests of the presence or absence of these multiproduct properties. This paper summarizes the results of work presented elsewhere in detail (see Baumol 1977 and Panzar and Willig 1977) describing the basic multiproduct attributes needed to generalize the concept of scale economies; we then report results of some empirical studies investigating the costs of producing a number of scientific journals which seem to confirm the presence of all of the generalized scale-economies properties, at least for the range of output levels for which data were available.[1]

I. Multiproduct Attributes Related to Scale Economies

To make it possible to follow our empirical analysis it is necessary to describe three rather technical concepts which are discussed in detail elsewhere (Baumol 1977):

Definition 1.—A cost function is said to be *subadditive* if (with given input prices) one firm can produce a given output vector, y^*, more cheaply than it can be produced by any combination of $m \geq$ two firms, each of them with the same cost function, $C(y)$. More formally, any $C(y)$ is subadditive at the output vector y^* if for all sets of output vectors y^i such that $\sum y^i = y^*$, $\sum C(y^i) > C(y^*)$. Subadditivity is obviously the basic criterion which must be satisfied if cost savings are to be achieved by combining firms (publishers).

Definition 2. Declining ray average cost.—While we cannot define average costs for a multiproduct firm since we cannot construct an unambiguous index of output for a set of heterogeneous products, we can determine how costs relate to output scale when outputs increase proportionately (i.e., as we move away from the origin along any ray in output space). More formally, for the ray through some given output, y', which we arbitrarily set equal to unity, we can define ray average cost as $RAC = C(ky')/k$. Then we have the following criterion of declining ray average cost along the ray through y': RAC is strictly declining if $C(ky')/k > C(vy')/v$ for $v > k$.

Declining ray average cost clearly corresponds to one variant of what we frequently mean by scale economies in the single-product case. However, it does not get us everything we require for a multiproduct analysis where it is not enough to know only whether it is cheaper (in terms of cost per unit) to produce more of a given product or of a given bundle of products in fixed proportions. It is also essential to determine whether it is more economical to produce several products (several different journals) *together*. Obviously, it will do two publishers no good to combine if ray average costs are declining but there is no saving in the simultaneous production of the journals of each publisher (when there is no expansion in the number of copies of *each* journal supplied).

---

[1] For earlier estimates of scale economies in journal production see Berg (1970, chap. 2).

Accordingly, we have the following concept which is designed to capture this sort of interproduct complementarity:

Definition 3. Trans-ray convexity.—If at a point, y, in output space there exists at least one negatively sloping (trans-ray) cross section such that costs are no higher toward the interior of the cross section than they are toward its edges (it is cheaper to produce mixtures of outputs than to produce outputs entirely or nearly in isolation), we say the cost function is trans-ray convex at y. More formally, a cost function $C(y)$ is trans-ray convex at y if for the output vector $y = (y_1, \ldots, y_n)$ there exists *any* set of positive constants (product "prices") $w_1, \ldots, w_n$, such that for every two output vectors $y^a = (y_{1a}, \ldots, y_{na})$, $y^b = (y_{1b}, \ldots, y_{nb})$ satisfying $\sum w_i y_{ia} = \sum w_i y_{ib} = \sum w_i y_i$, we have $C[ky^a + (1 - k)y^b] \leq kC(y^a) + (1 - k)C(y^b)$ for any k, $0 > k > 1$.

Unfortunately, there is no way we can determine from information in the neighborhood of output vector y^* alone whether cost is subadditive at y^*. The reason is straightforward. Subadditivity means that given an industry output vector y^* initially produced by a single firm, it must *never* be cheaper to break off any small (or large) output vector Δy^* and to produce Δy^* by itself. That is, it must never be cheaper to go from a single producer turning out y^* to two producers turning out $y^* - \Delta y^*$ and Δy^*, respectively. But to test for this we must know $C(y^*)$, $C(y^* - \Delta y^*)$, and $C(\Delta y^*)$ for *all* $\Delta y^* < y^*$, including Δy^* very small (i.e., very near the origin). Thus, the empirical investigator who is testing for subadditivity is faced with the difficult task of finding the shape of the cost function in the region of the origin, for which no data may be available.

Some help in this task may be obtained from

Proposition 1.—For any given y^*, strictly declining RAC for $y < y^*$ together with (nonstrict) trans-ray convexity along any one cross section through y^* are sufficient for subadditivity at y^* (for proof see Baumol 1977).

With these materials we can proceed directly to our empirical results.

II. An Estimate of Economies of Scope in Journal Production

Using cost and output data from a sample of nonprofit journal publishers (e.g., professional societies and university presses), we were able to test directly for the presence of economies of scope in journal publication.

(Following Panzar and Willig [1977], economies of scope are said to exist when it is cheaper to produce the outputs in combination rather than separately.) These data came from publishers whose numbers of journals ranged from one to 36 in a variety of disciplines. We have figures for three years, 1969, 1971, and 1973.[2] Fifty-six journals provided the necessary data and therefore comprised our sample.

The cost functions estimated from this sample did turn out to exhibit economies of scope. This was true both of additive and multiplicative forms of the cost functions. In each case costs increased more slowly with increases in circulation than with increases in pages, and in both cases costs per journal declined as the number of journals per publisher increased. The estimated equations are[3]

$$\begin{array}{llllll} TC = & -3.44 & + 2.81Ds & + 0.584C & + 0.969P & - 0.155J \\ & (-4.65) & (3.11) & (11.16) & (20.19) & (-4.68) \end{array} \tag{1}$$

$$R^2 = .8240,\ F = 190.75,\ N = 168;$$

$$\begin{array}{lll} \log TC = -0.735 & + 0.171Ds & + 0.588 \log C \\ \quad (-8.29) & (2.35) & (15.96) \\ + 0.793 \log P & - 0.062 \log J & \\ \quad (15.96) & (-2.52) & \end{array} \tag{2}$$

$$R^2 = .8524,\ F = 235.41,\ N = 168;$$

where TC = total cost per journal (in \$10,000s); C = circulation (in 1,000s); P = pages (in 100s); J = number of journals published by this publisher, and Ds = 1.0 if the journal is scientific or technical, 0.0 if the journal is not scientific or technical.

The logarithmic form (eq. [2]) provides direct information about the magnitude of the economies of scope. For the same number of pages and the same circulation and field, the costs per journal of a publisher with 36 journals were 80 percent of those of a publisher who only published one journal (i.e., $36^{-.062} = 0.8008$).

III. Subadditivity and Interproduct Complementarity in Journal Supply

Our next step was to design a specification for a cost function that would enable us to measure directly the magnitude of interproduct complementarity but would not prejudge the existence of such complementarities. One specification that was tried was

---

[2] The data sources are described in Appendix A.

[3] The regression program used an error-components model to pool the cross-section and time-series data. This procedure assumes that the error term consists of three independent, random components—one from the cross sections, one from the time series, and an overall component variable in both. The data here satisfied these assumptions as well as requirements of having the components independent of the other variables. Also, more complex deflators did not perform as well as the dummy variables used in the first iteration. The t-statistics shown below each coefficient and the other summary statistics are not exact in the finite-sample case (see Wallace and Hussain 1969).

$$TC = b_0 + b_1Y_1 + b_2Y_2 + b_3Y_1Y_2 \tag{3}$$

where TC is the total cost of the firm and Y_1 and Y_2 are two outputs or classes of outputs. Depending on the estimated values of the coefficients, b_i, "average" costs may increase or decline along any ray from the origin. Similarly, the trans-ray properties of the function depend upon the estimates of b_3. By construction of (3), the marginal cost of either output is constant for all levels of *that* output. To test for declining ray average cost, we note that along any ray $Y_2 = vY_1$, $v > 0$ so that (3) becomes $TC = b_0 + (b_1 + vb_2)Y_1 + vb_3Y_1^2$; average cost is $b_0/Y_1 + (b_1 + vb_2) + vb_3Y_1$, which is strictly declining if b_0 turns out to be positive and b_3 is negative. The condition $(\partial^2 TC)/(\partial Y_1 \partial Y_2) = b_3 < 0$ is necessary and sufficient for the presence of trans-ray convexity since along a (negatively sloping) linear cross section we must have $Y_2 = -wY_1 + k$, $w > 0$.[4] When substituted into (3) this gives us $TC = b_0 + kb_2 + (b_1 - wb_2 + kb_3)Y_1 - wb_3Y_1^2$ whose second derivative is $-2wb_3$, which is positive if and only if $b_3 < 0$. Therefore, by proposition 1 the two conditions $b_3 < 0$ and $b_0 > 0$ *are sufficient for subadditivity if (3) is an adequate* representation of the cost function for all points closer to the origin than those actually observed.

A journal publisher who is engaged in the publication both of journals containing original material and journals containing translations of works first published in another language provided our next set of data. This set was particularly interesting as the dependent variable is the annual publishing expenses of the firm—no artificial allocation of costs among the journals was involved. However, we experienced two types of problems in our attempts to estimate the coefficients of cost functions of the type in equation (3). First, we had only 9 years of data, causing a shortage of degrees of freedom. This and the correlation between the cross-product term and the Y_1 and Y_2 resulted in low statistical significance. Second, we had difficulty in characterizing the outputs. We used circulation, pages, or impressions (pages × circulation) as explanatory variables, but could not use them in combination because of the shortage of degrees of freedom. The most promising regression result was

$$\begin{aligned} TC = \underset{(0.27)}{-17.7} + \underset{(0.24)}{6.93P_1} + \underset{(0.34)}{13.1P_2} - \underset{(0.27)}{4.50P_1P_2} \end{aligned} \tag{4}$$

$$R^2 = .5174,\ F = 1.787,\ N = 9.$$

Obviously, the low significance prohibits the drawing of any conclusions from equation (4).

---

[4] Note that the condition $(\partial^2 TC/\partial Y_1 \partial Y_2) = b_3 < 0$ is also often interpreted as the condition of interproduct complementarity.

In an effort to improve the level of statistical significance of the regressions and to capture more fully the economics of the publishing process, we undertook two further steps. First, we designed a scalar measure of the outputs of the journal publishing process,[5] and second, we generalized our cost function. This more general cost function takes the form

$$TC = b_0 + b_1S_1 + b_2S_2 + b_3(S_1S_2)^\gamma \tag{5}$$

where S_1 and S_2 are the scalar representations of the outputs and the b_i and γ are parameters whose values were to be estimated.

Relationship (5), as we will see, turns out to be far richer than (3) both as an instrument for empirical estimation and as a basis for the design of practical policy. That is because (5) may or may not exhibit monotonic behavior of the ray average costs. Depending on the values of the parameters derived from the data, the ray average cost curves may (or may not) turn out to be U-shaped. If they are, there will be a locus of points of minimum ray average costs, one such point lying on each ray in output space. That locus will, in general, determine whether costs can be saved by amalgamation of some firms, whether they can be saved by subdividing some large firms into smaller ones, or whether neither of these steps is called for.

However, functional form (5) introduces several complications. The first is the nonlinear nature of the cross-product term which requires the use of an appropriate nonlinear estimation technique.[6] The second is the fact that the criterion of declining ray average costs no longer depends solely on the signs of b_0 and b_3 alone. There are now two cases. In case 1, where $1 \geq \gamma \geq 1/2$, ray average costs must decline if $b_0 > 0$ and $b_3 < 0$ as before. But in case 2, where $1/2 > \gamma > 0$, ray average costs will decline if and only if

$$\left[\frac{b_0}{(2\gamma - 1)b_3}\right]^{1/\gamma} > S_1S_2.^7 \tag{6}$$

Thus, the boundary of the region of declining ray average costs, which is obtained by replacing the > in (6) by an equals sign, is clearly a rectangular hyperbola in S_1, S_2 space.

---

[5] The derivation of these scalars is described in Appendix B.

[6] We employed the nonlinear least-squares (LSQ) procedure of the Harvard University version of Times-series Processor (TSP).

[7] *Proof*: We have $TC = b_0 + b_1S_1 + b_2S_2 + b_3(S_1S_2)^\gamma$ where $1 > \gamma > 0$. To test for the behavior of cost along a ray, set $S_2 = vS_1$. Then $TC = b_0 + b_1S_1 + vb_2S_1 + v^\gamma b_3S_1^{2\gamma}$, $AC = b_1 + vb_2 + b_0/S_1 + v^\gamma b_3S_1^{2\gamma-1}$, and $dAC/dS_1 = -b_0/S_1^2 + (2\gamma - 1)v^\gamma \times b_3S_1^{2\gamma-2}$. It is clear that if $b_0 > 0$, $b_3 < 0$, $1 > \gamma \geq 1/2$, the preceding derivative must be negative, i.e., ray average costs must decline everywhere. However, if $b_0 > 0$, $b_3 < 0$, $1/2 > \gamma > 0$, then $dAC/dS_1 = -b_0/S_1^2 + (2\gamma - 1)v^\gamma b_3S_1^{2\gamma-2} < 0$ if $b_0 > (2\gamma - 1)v^\gamma \times b_3S_1^{2\gamma}$, or $[b_0/(2\gamma - 1)b_3]^{1/\gamma} > S_1S_2$.

It follows from proposition 1 that if $b_0 > 0, b_3 < 0, 0 < \gamma < 1/2$, then the region of subadditivity must include all combinations of S_1, S_2 that satisfy (6). However, since the requirements of proposition 1 are sufficient but not necessary for subadditivity, the region of subadditivity may extend beyond this. We can, however, prove that the outer bound of the region of subadditivity must lie in the region $[b_0/(2\gamma - 1)b_3]^{1/\gamma} \leq S_1S_2 \leq 4[b_0/(2\gamma - 1)b_3]^{1/\gamma}$.[8]

All of this means that in the case $b_0 > 0, b_3 < 0, 0 < \gamma < 1/2$ there is a range of outputs within which savings can be achieved by amalgamation of firms, but, that once amalgamation moves the firm beyond the region of subadditivity, as indicated by the preceding inequalities, the process of concentration can add to total costs rather than decrease them.

In (5) the conditions $b_3 < 0$ and $1 \geq \gamma > 0$ are sufficient for trans-ray convexity since along a linear cross section, $S_2 = -wS_1 + k$, the only nonlinear term in (5) becomes $b_3(kS_1 - wS_1^2)^\gamma$, whose second derivative is $(\gamma^2 - \gamma)b_3(kS_1 - wS_1^2)^{\gamma-2}(k - 2wS_1)^2 - 2w\gamma b_3(kS_1 - wS_1^2)^{\gamma-1} > 0$ if $b_3 < 0, 1 \geq \gamma > 0$. Using the same data as before to obtain parameter values for (5) we obtained the cost function

$$\begin{aligned} TC = {} & \underset{(2.88)}{582.8} + \underset{(2.73)}{2.38S_1} + \underset{(2.82)}{8.00S_2} - \underset{(-2.84)}{345.7(S_1S_2)^{0.13}} \qquad (7) \\ R^2 = {} & .8012,\ F = 6.72,\ N = 9 \end{aligned}$$

where TC = annual costs (in \$ millions).

The estimation procedure was sensitive to the choice of initial values. The t-statistics (shown in parentheses below the coefficients) are approximate. Since both the coefficient and exponent of the cross-product term (b_3 and γ, respectively) are important for the existence of trans-ray convexity, a likelihood ratio test (following Goldfeld and Quandt 1972) was utilized to test the hypothesis that b_3 or γ or both equal zero.[9]

Let λ be the likelihood ratio

$$\lambda = \frac{L(S \mid \overline{b_0}, \overline{b_1}, \overline{b_2})}{L(S \mid \hat{b}_0, \hat{b}_1, \hat{b}_2, \hat{b}_3, \hat{\gamma})} \qquad (8)$$

where the $\hat{b}$ and $\hat{\gamma}$ are obtained from equation (6) and the b are from the (conditional) least-squares equation without the cross-product term. Therefore, if SSR is the sum of squares of the residuals from equation (6) and SSR_c is the sum of squares of the residuals from the conditional regression, we have

$$-2 \log \lambda = n \log (\text{SSR}_c/n) - n \log (\text{SSR}/n). \qquad (9)$$

---

[8] For proof, see Baumol and Fisher, in press. The argument, intuitively, is that any S_1, S_2 which equals $[b_0/(2\gamma - 1)b_3]^{1/\gamma}$ lies on the point of minimum ray average cost along the ray through that point. Hence, any firm producing at least $2S_1$, $2S_2$ can be replaced by two smaller firms with some saving in total cost.

[9] Goldfeld and Quandt 1972, pp. 72–74.

Although the low number of observations causes concern, under the appropriate assumptions $-2 \log \lambda$ follows a χ^2 distribution with 2 df.

For equation (7), $-2 \log \lambda = 8.65$, and we can reject the hypothesis that b_3 and/or γ equal zero at the 95 percent significance level.[10] This provides a confirmation of the presence of trans-ray convexity. To test for the presence of declining ray average costs we use inequality (6). Substituting in the estimated values of b_0, b_3, and γ, we obtain

$$\{582.8/[(2 \times 0.13) - 1][-345.7]\}^{1/0.13} = 563.2 \text{ (approx.)} > S_1S_2,$$

that is, ray average costs will decline for all values of S_1, S_2 which lie between the origin and the rectangular hyperbola $S_1S_2 = 563.2$.

In fact the observations in the case we are now considering all lay very close to the locus of points of minimum ray average costs, $563.2 = S_1S_2$ corresponding to (6). That is, it is the locus of points, along each ray in output space, at which the U-shaped cross section of our cost surface (7) reaches its minimum. Specifically, two-thirds of our observations lie just inside that locus, in the direction of the origin, and the remaining points lie just outside that locus.

We know by trans-ray convexity that all points between the locus of minimum RAC and the origin lie in the region of subadditivity, and, by continuity, this is probably true for points just outside that locus. Thus, for firms of the size in our range of observations it seems very likely that no savings can be achieved by subdivision into smaller enterprises. But will it pay to amalgamate two such firms? Almost certainly not, for if two such firms combine their output vector will lie well outside the locus of minimum ray average cost and therefore beyond the known region of subadditivity. We can only be sure that it pays to amalgamate two firms if their output vectors both lie much closer to the origin, indeed inside a curve half as far from the origin as the locus of minimum ray average costs, because only then can the vector sum of these two output points be guaranteed to lie inside the region subadditivity.

We conclude that, whatever the reasons, the firm observed in this analysis seems to have grown to a size that is roughly consistent with cost efficiency.

IV. Complementarity between Publication and Membership Activities

We have already noted that it is important for a test of subadditivity that the estimated values of the parameters be valid near the origin. In the analysis of the preceding section the observed values of the independent variables unfortunately encompassed a limited range in output space; S_1 varied between 36 and 46 while S_2 varied from 11 to 14. We therefore turn, next, to a case in which the observed values of the variables occupied a wider range of values.

---

[10] A maximum-likelihood approach can also be employed to test the significance of all of the estimated parameters (ibid., pp. 68–72).

For many scholarly and professional societies, publishing constitutes a major part of their overall activity. We obtained data from an organization which has affiliated with it a large number of societies, each of which publishes a journal as well as providing other services and activities for its members. The data covered 14 years of operation for 25 of the affiliated societies. Using these data and a specification similar to equation (5), we estimated a cost function that would permit us to detect subadditivity if it were present.

The estimated function is

$$\begin{array}{lllll} TC = 153.4 + & 1.86M + & 0.680P - & 13.17(MP)^{0.1} & \\ \quad\ (6.63) & (20.14) & (12.72) & \quad(-7.81) & \qquad (10) \\ R^2 = .7937, & F = 443.8, & N = 350 & & \end{array}$$

where TC = annual society costs (in \$10,000s), P = pages (in 100s) and M = membership (in 1,000s). The likelihood ratio, calculated in the same manner as before, gives $-2 \log \lambda = 59.38$, and we can again reject the hypothesis that the cross-product term is not significant.

Here the data are all within the boundary delimited by (6), which in this case takes the form $S_1S_2 < 4.28 \times 10^{11}$! Since the maximal value of S_1S_2 in the observed data is 540, all our tests of the presence of transray convexity and subadditivity are satisfied within the range of outputs considered, if the relationship (10) holds up to the origin. This time the range of the observations was much greater than in our previous calculation. The number of pages varied from 20 to 3,000 per year and membership ranged from 600 to 18,000, giving us observations that can be considered to fall very close to the axes and the origin.

In this case we seem to be able to draw the conclusion that the organizations represented would, at least in many cases, benefit from amalgamation since the zone of subadditivity extends throughout the region about which we have information. However, we are less certain of the implications of this cost function because there is necessarily a high correlation between membership and circulation and because we were forced to use number of pages alone to measure the publishing activity of the societies.

V. Concluding Comment

The preceding results show that the special concepts needed to deal on a theoretical level with scale economies phenomena in multiproduct firms and industries are also amenable to empirical investigation. Specifically, despite its nonlocal character, it seems to be possible to draw some inferences about the presence or absence of subadditivity. Of course, much more work will still have to be done on the testing of that fundamental concept. But we feel that for a first venture into this area our study has produced fairly strong conclusions and fairly strong evidence of the tractability of the subject.

On an applied level the analysis seems to show clearly that these phenomena are all present in the case of journal publication. They suggest that at current scales of operations and numbers of journals, savings are sometimes achievable through amalgamation of publishers. Given the financial pressures that beset them, that may provide an important contribution toward the alleviation of their problems. Of course, amalgamation need not be the only way in which such savings can be achieved—for example, one publisher may achieve them by expanding into the publication of a number of new journals and, with his reduced costs, run the other publishers out of business, provided of course that these new journals matched the specialization and quality of the old ones with which they compete.

But this is all conjecture. Data for other publishers may turn out to indicate that the subadditivity results do not hold throughout the industry. The evidence suggests, moreover, that they hold only for a fairly limited range of outputs. Indeed, it would hardly seem plasuible that journal publication is a hitherto undiscovered natural monopoly.

In short, we claim merely to have indicated the workability of an empirical investigation of the issue of subadditivity which must underlie any systematic analysis of the cost effectiveness of the organization of an industry.

Appendix A

Data Sources

Equations (1) and (2)

The data for these estimations came from a survey conducted by B. Fry of nonprofit publishers of scientific journals (see Fry and White 1975). The data were supplied to us with the identities of the publishers removed, making it impossible to know the nature and degree of hidden subsidies (e.g., donated editors' time or space) or the effects of different accounting practices, etc.

Equations (3)–(10)

The data for these estimations were collected from two professional societies with the assistance of F. Machlup et al. Here, every effort was made to include all relevant costs. However, the true values of the editors' contributions of time and effort were either understated or omitted entirely.

Appendix B

Scalar Measures of Publication Output

It is obvious from the preceding discussion that it is very useful to have a scalar index that describes effectively the output of the journal publishing process. In addition to number of pages published (and measures correlated with number of pages, such as words, articles, etc.) and circulation (or press run), one possibility that has often been suggested as a scalar output measure is impressions—pages × circulation. The logic underlying this choice is that the number of impressions is the number of times the "press hits the page" (or its modern technological equivalent).

We have statistics to measure the adequacy of these candidate indices. Tables B1 and B2 present the coefficients of determination (R^2) obtained from a simple regression of cost on each of these indices. The results both for a linear specification ($TC = a + bX$) and for a multiplicative specification ($TC = aX^b$) are presented.[11]

The usefulness of a composite variable such as impressions can also be measured in another way. The use of number of impressions in a simple multiplicative regression is equivalent to restricting the exponents of pages and circulation to equal values in a multiple regression with the form $TC = aIm^b = b_0P^{b_1}C^{b_2}$ if

TABLE B1

FRACTION OF VARIATION IN OUTPUT (R^2) ACCOUNTED FOR BY SINGLE EXPLANATORY VARIABLES (Linear Specification)

	DATA SET			
VARIABLE	Fry Sample ($N = 56$)	Fry Multijournal Publ. Sample ($N = 84$)	Publisher 1 ($N = 54$)	Publisher 2 ($N = 90$)
Pages.......... ..	.69	.78	96	.89
Circulation..	.27	.22	.15	.64
Impressions	.81	.75	.95	.91
Scalar	.89	.91	.97	.90

TABLE B2

FRACTION OF VARIATION IN OUTPUT (R^2) ACCOUNTED FOR BY SINGLE EXPLANATORY VARIABLES (Multiplicative Specification)

	DATA SET			
VARIABLE	Fry Sample ($N = 56$)	Fry Multijournal Publ. Sample ($N = 84$)	Publisher 1 ($N = 54$)	Publisher 2 ($N = 90$)
Pages.	.60	.65	.91	.91
Circulation.......	.59	.51	.30	.77
Impressions	.84	.84	.97	.94
Scalar	.85	.86	.98	.94

---

[11] As is common practice, the multiplicative specifications were estimated by a linear regression after taking logarithms of both sides.

TABLE B3

TESTS FOR EQUALITY OF EXPONENTS OF PAGES AND CIRCULATION VARIABLES

Data Set (N)	b_1 (Pages)	b_2 (Circulation)	t (Calculated Test Statistic)
Fry (168)	.814 (.0474)	.559 (.9367)	3.59
Fry (multijournal publ.) (84)	.825 (.0587)	.534 (.0490)	3.28
Publisher 1 (54)	.718 (.0162)	.472 (.0326)	6.07
Publisher 2 (90)	.752 (.0482)	.494 (.0751)	2.19*

NOTE.—Standard errors are below each exponent in parentheses. The numbers under the data sets are the sample sizes.

* Cannot reject hypothesis of equality at 99% level

$Im = P \cdot C$ and $b = b_1 = b_2$. We can then use the estimates of b_1 and b_2 from the multiple regressions to test for their equality.[12] These values and the results of the tests are reported in table B3.

These results turn out to have rather interesting features. We can reject the hypothesis of equality of the exponents in all four of the cases at the 95 percent significance level and in three of the cases at the 99 percent level. Even with the limited range of the ratio of the two exponents (from 1.45 to 1.55), there is an indication that a scalar of the form $S = P^{b_1}C^{b_2}$ with $b_1 > b_2$ may account for a greater portion of the variation in costs than number of pages, circulation, or impressions. The proportions of the variation accounted for by each of these indices are shown in tables 1 and 2. The scalar S was, consequently, the index used in Section III.

References

Baumol, W. J. "On the Proper Cost Test for Natural Monopoly in a Multi-Product Industry." *A.E.R.*, vol. 67 (December 1977).

Baumol, W. J., and Fischer, D. "Cost Minimizing Number of Firms and Determination of Industry Structure." *Q.J.E.* (1977), in press.

Berg, S. "Structure, Behavior, and Performance in the Scientific Journal Market." Ph.D. dissertation, Yale Univ., 1970.

Fry, B., and White, H. J. "Economics and Interaction of the Publisher-Library Relationship in the Production and Use of Scholarly and Research Journals." Final report, National Science Foundation Grant GN-4398. Springfield, Va.: Nat. Tech. Information Service, 1975.

Goldfeld, S. M., and Quandt, R. C. *Nonlinear Methods in Econometrics.* Amsterdam: North-Holland, 1972.

Panzar, J. C., and Willig, R. D. "Economies of Scale and Economies of Scope in Multi-Output Production." Bell Laboratories discussion paper 33, Bell Laboratories, Holmdel, N.J., August 1977.

Wallace, T. D., and Hussain, A. "The Use of Error Components Models in Combining Cross-Section and Times Series Data." *Econometrica* 37 (January 1969): 55–72.

---

[12] The test statistic is $t = C'b/\sqrt{C' \sum C}$ where $C' = (0, 1, -1)$, $B' = (b_0, b_1, b_2)$, and $\sum$ is the variance-covariance matrix of the estimated coefficients.

The IS-LM Model of Macroeconomic Equilibrium and the Monetarist Controversy

Neil Bruce

Queen's University

This paper highlights the issue of whether government bonds are net wealth as an important one in the monetarist controversy. In particular it is shown that the standard IS–LM model and its conclusions arise as a special case of a more general macroeconomic model which results when government bonds are fully perceived as collective wealth. When, however, the tax liabilities required to service that debt are fully capitalized, a model with monetarist conclusions emerges.

The resolution of theoretical differences between monetarists and income expenditure theorists[1] has long been impeded by the lack of a common theoretical framework. Recently, however, some leading monetarists have shown an increased willingness to express their ideas in terms of the IS-LM model of macroeconomic equilibrium which is central to the income expenditure view.[2] But unlike income expenditure theorists, monetarists have refused to conclude that the crucial issues in determining the relative effectiveness of monetary and fiscal policy relate to the slopes of the IS and LM curves.[3] In this paper it will be argued that the IS-LM

I am indebted to R. Boadway, J. Frenkel, R. Lipsey, L. Paquin, D. Purvis, G. R. Sparks and an anonymous referee for numerous comments, criticisms, suggestions, and corrections on earlier drafts. I remain responsible for any remaining errors and omissions.

[1] By income expenditure theorists I mean the self-professed "eclectics" who accept the conclusions regarding macroeconomic policy which follow from the textbook IS-LM analysis.

[2] For example, Milton Friedman has declared the IS-LM equations "accepted alike by adherents of the quantity theory and of the income expenditure theory" (1971, p. 31). It should be noted, however, that he precedes this statement with a warning that algebraic equations can conceal "fundamental differences." See further discussion in Tobin (1972) and Friedman (1972).

[3] This refusal has resulted in an article by Blinder and Solow (1973) who attempt to resolve monetarist income expenditure theory differences in terms of a "third level" of crowding out resulting from wealth arguments in the portfolio balance and saving functions. I have abstracted from these direct wealth effects so as to clarify the central argument of this paper.

[*Journal of Political Economy*, 1977, vol 85, no. 5]
© 1977 by The University of Chicago All rights reserved

model is deficient from a monetarist point of view in that it fails to fully include the flow supply of newly issued money in its flow equilibrium conditions. An examination of the circumstances under which such an omission is valid reveals that the question of whether or not interest bearing government debt is perceived as collective net wealth by the private sector may be an important one in the monetarist controversy. In particular, it will be shown that when the taxes required to service the interest bearing government debt are fully capitalized by the private sector the IS-LM model should be modified in a way which yields conclusions regarding the relative effectiveness of monetary and fiscal policy similar to those held by monetarists.

Before proceeding, it would be indicated that I am not addressing the difficult question of whether changes in aggregate demand cause price level or real output changes, hence I will assume that the price level is constant and that all changes in nominal income are caused by changes in output. I also assume that all money is of the "outside" variety so that the existence of financial intermediaries and their effect on the liquidity ratio of private wealth can be ignored.

I. Portfolio Balance and the IS-LM Model

A major reason why the IS-LM model is so useful and illuminating lies in the basic assumption of the model, which states that the conditions of stock and flow equilibrium can be analytically separated. In Tobin's words (1969, p. 16), "a key . . . assumption . . . is that spending decisions and portfolio decisions are independent-- specifically that decisions about the accumulation of wealth are separable from decisions about its allocation. As savers, people decide how much to add to their wealth; as portfolio managers, they decide how to distribute among available assets and debts the net worth they already have." This assumption is enormously simplifying and will be maintained in this paper, but at some cost. It should be recognized that the conditions under which such separation is valid are very stringent and that the separability assumption itself is probably a factor in the monetarist controversy.[4] The importance of the assumption is that the stock equilibrium conditions (the LM curve) and the flow equilibrium conditions (the IS curve) can be derived separately. Although the central argument of this paper is that the main differences between monetarists and income expenditure theorists lie in the determination of flow equilibrium, it is useful at this point to clarify the exact meaning of the portfolio balance conditions as given by the LM curve.

---

[4] The obvious condition which is necessary for portfolio choice decisions to be made independently from the savings-consumption decision is that the borrowing and lending rates facing the wealth holders be equal. Otherwise the interest rate on nonmonetary assets will depend on whether the wealth holder is saving or dis-saving, hence his portfolio balance conditions cannot be specified independently of this fact. Also, when uncertainty is present, perfect (costless) insurance markets are required (see Drèze and Modigliani [1972]). The implications for monetary and fiscal policy when households' choices cannot be separated are examined by Purvis (1975).

The LM relation is well developed in terms of portfolio theory, particularly due to the work of James Tobin (1969).[5] The simplest possible LM relation is assumed in this paper, that which holds for an economy with three assets—money, interest-bearing government bonds, and equities—where the latter two are perfect substitutes. The nominal values of these assets are denoted M, B, and E, respectively. Government debt consists of b consols, each paying a flow return of one nominal dollar and having a market price of P_b dollars. Following Patinkin (1965, p. 289), I let $0 \leq k \leq 1$ be the constant proportion of the value of each government bond perceived as net collective wealth by the private sector. Although each bond is treated as wealth by a single individual, collectively bonds may not be so treated because the private sector may capitalize some part of the future tax liabilities needed to service that debt.[6] The difference between the value of the bond and the capitalized value of its future tax liabilities divided by value of the bond equals the constant k. Equities consist of e shares, each paying a flow return of one real dollar (or P nominal dollars, where P is an aggregate price level) and having a market price of P_e dollars. The value of equities is allowed to diverge from the displacement value of the physical capital it represents.[7] Finally, the pecuniary return on money is institutionally set at zero, but monetary assets yield liquidity services to the holder. The outside money stock held by the private sector is assumed to be treated as collective wealth, since the nonpecuniary services which give money its positive value do not entail future tax liabilities.

The portfolio balance conditions are represented by the LM curve, which is the locus of interest rates on nonmonetary assets (r) and levels of real output (y) such that the private sector is content with the *composition* of the aggregate portfolio. In the simple economy described, there are effectively only two assets since bonds and equities are perfect substitutes; thus Walras's Law of stocks permits the portfolio balance conditions to be represented by the single, familiar equation:

---

[5] See also the seminal paper by Metzler (1951).

[6] The assumption that future tax liabilities are capitalized is an important one. It is reasonable to the extent that taxes are imposed on the income streams of marketable assets such as would be the case for corporation income taxes (see Mundell [1971, p. 11]). In fact, some portion of these taxes fall on wage income and even if perceived they would not be capitalized at the market rate of interest because of the inability to borrow against human capital. As a result the future tax liabilities should not be treated as equivalent to the marketable components of private sector wealth. To properly deal with this problem, it would be necessary to introduce a nonmarketable asset, such as human capital, from which tax liabilities can be subtracted. This would necessarily require dropping the IS-LM framework since the validity of the separability assumption would be questionable under such conditions. In this paper it can be assumed that tax liabilities on wage income are capitalized at some "psychic" discount rate exceeding the market rate, thus implying $k > 0$.

[7] That is, the Tobin "q" (1969, p. 21) is not constrained to unity. The usual story behind this assumption involves a rising supply price of investment goods, e.g., Witte (1963), or a Penrose effect due to adjustment costs as used by Frenkel and Rodriguez (1975).

$$M^s = P \cdot l(y, r); \qquad l_1 > 0, l_2 < 0, \tag{1}$$

where M^s is the nominal stock of money balances. The demand for real money balances is denoted $l(\cdot)$ where this function could include the real value of private wealth as suggested by Silber (1970, p. 464) and pursued by Blinder and Solow (1973). This wealth dependency is not directly relevant to this paper and is therefore excluded.

It bears repeating that equation (1) represents equilibrium as to the composition of private sector wealth between monetary and nonmonetary assets and does not imply that the actual and desired levels of private wealth are equal. In fact, the portfolio balance conditions must obtain before a wealth aggregate can be evaluated. That is, equation (1) determines a vector of nominal asset prices:[8]

$$\begin{aligned} P_m &\equiv 1 \\ P_b &= \frac{1}{r} = \frac{1}{\phi(y, M^s/P)} \\ P_e &= \frac{P}{r} = \frac{P}{\phi(y, M^s/P)}, \end{aligned} \tag{2}$$

where $\phi(\cdot)$ is found by solving (1) for r. Using equations (2), the real value of existing private wealth can then be evaluated as:

$$\begin{aligned} w = W/p &= \frac{M^s + k \cdot p_b \cdot b + p_e \cdot e}{P} \\ &= \frac{M^s}{P} + \frac{k \cdot b + P \cdot e}{P \cdot \phi(y, M/P)}. \end{aligned} \tag{3}$$

From (3) it is clear that given values of M^s, b, e, y, and P, the portfolio balance conditions determine a value of w. The obvious extension of monetarist reasoning in this context is to specify a real wealth demand function which, in conjunction with existing w, determines whether intended expenditure is greater, less than, or equal to given output y. It is at this juncture that differences between the monetarist and income expenditure theories arise.

---

[8] Constant relative risk premiums and expected capital gains and losses could be added to the r discounting each asset's flow return, but these considerations do not alter the main argument and are excluded.

II. The Flow Equilibrium Conditions and the Flow Demand for New Wealth

The flow equilibrium conditions in the standard IS-LM model are obtained by equating the sum of possible uses of personal income ($c + s + t$) and national product plus transfers ($y + v = c + i + g + v$) where y, c, s, t, i, g, and v denote the real values of national product, consumption expenditure, saving, personal taxes, investment expenditure, government expenditure on current goods and services, and transfers, respectively. By specifying behavioral functions for saving and investment, the familiar IS equation is obtained as:

$$s(y_d) = i(r) + g - t', \tag{4}$$

where $t' = t - v$ denotes personal taxes net of transfers and $y_d = y - t'$ is disposable income.

Ex post, equation (4) is an accounting identity and bereft of theoretical interpretation; ex ante, the variables in (4) are also intended values and equation (4) represents an equilibrium condition. The main point I wish to make is that as an equilibrium condition, equation (4) and the definition of saving therein is too narrow because certain changes in net private wealth should be included which are presently being ignored. For example, Shell, Sidrauski, and Stiglitz (1969) have argued that the private sector's purchasing power (which is allocated to c, s, and t) should include capital gains and losses on existing assets. Thus intended savings (the desired rate of accumulation of private sector wealth) must be defined to include such capital value changes while the right-hand side of (4) (the flow supply of private sector wealth) should include them as a distinct element. Since I will focus on an economy where commodity and asset prices are constant, capital gains and losses will not be incorporated into the flow conditions in this paper, although it would be necessary to do so if an inflationary context were to be considered. The important point of this paper is that a similar consideration arises when the tax liabilities of interest bearing government debt are capitalized and that equation (4) does not include the changes in the capitalized value of these tax liabilities which result when the size of the interest bearing government debt is altered by bond financed government deficits or by open market operations. For the former, each new dollar of government debt is perceived as k dollars of collective private sector wealth, while the expenditure or transfer it financed was fully counted as personal income. As a result, a further wealth change of $k - 1$ times the value of the bond financed government deficit should be included in the definition of desired saving and in the right-hand side of (4). Also, since open market operations involve the exchange of money, which is assumed to be fully perceived as net wealth, for bonds which are not, a further change in private sector wealth of $1 - k$ times the flow supply of money resulting from open market operations should be included.

Combining the above, we can obtain an expanded measure of the private sector's disposable income, available for consumption and changing its net wealth position (saving) as:

$$y_d^a = y - t' - (1 - k)\left(g - t' - \frac{1}{P}\frac{dM_g}{d\tau}\right) + (1 - k)\frac{1}{P}\frac{dM_b}{d\tau} \qquad (5)$$

where $(1/P)(dM_g/d\tau)$ is the flow supply of new money issue financing the government deficit (therefore $g - t' - (1/P)(dM_g/d\tau)$ is the deficit financed by new bond issue) and $(1/P)(dM_b/d\tau)$ is the flow supply of new money issue resulting from open market operations and is equal to the value of bonds exchanged.[9] Letting $dM/d\tau = (dM_g/d\tau) + (dM_b/d\tau)$, we can rewrite (5) as:

$$y_d^a = y - t' - (1 - k)\left(g - t' - \frac{1}{P}\frac{dM}{d\tau}\right). \qquad (5')$$

Letting $y = c + i + g$ and equating y_d^a to $c + s^a$, we obtain:

$$s^a(y_d^a) = i(r) + k(g - t') + (1 - k)\frac{1}{P}\frac{dM}{d\tau}$$

where $s^a(\cdot)$ is an expanded concept of saving which includes net wealth changes resulting from changes in capitalized future tax liabilities. It might be noted that $s^a(\cdot)$ differs from $s(\cdot)$ by exactly that expression which must be added to the right-hand side of (4) to obtain the right-hand side of (6). Hence as an ex post identity, equation (4) is intact. As an equilibrium condition, however, it is considerably altered since it is the augmented concept of saving on which the behavioral assumptions are made.[10]

To better understand these points, it should be realized that saving behavior should represent the desired rate of wealth accumulation by the private sector rather than being a "psychological" law or ultimate behavioral function in and of itself. Such a saving concept arises naturally when one takes a life-cycle or permanent-income view of the saving/consumption function. A simple model incorporating this view of saving will now be described and equation (6) will be found to also represent the equilibrium between the flow demand and the flow supply of private sector wealth.

---

[9] In a growth context, a continuous flow supply of new money balances is consistent with "neutral" monetary policy. In Section IV it is argued that for an economy with a stable price level, such neutrality requires that the rate of monetary expansion not exceed γ, the exogenous growth rate of full employment income. Changes in monetary policy involve changes in the nominal money stock which alters the flow supply of money balances entering the economy since $dM/d\tau = \gamma M(t)$.

[10] The saving function of eq. (4) implies that the accumulative behavior of individuals is only concerned with the equities, money and/or bonds financing $i + g - t'$. A broader view of private sector wealth requires a broader definition of saving.

Given that the actual real value of private sector wealth is as given by equation (3) using the portfolio balance conditions, the next step is to specify a demand function for real marketable wealth:

$$w^d = w[(y_d^a)^*]; \qquad w_1 > 0, \tag{7}$$

where (y_d^a) is the expected or permanent level of disposable income as defined in equation (5). It is reasonable to assume that a positive relation exists between w^d and $(y_d^a)^*$ since a larger permanent disposable income would allow the wealth holder to spread a higher level of consumption over his lifetime necessitating a larger stock of wealth in order to finance the higher consumption level in retirement and other nonearning periods.

In a static economy, $(y_d^a)^*$ and hence w^d would be constant, therefore nonzero intended saving could only arise for temporary stock adjustment reasons when desired and actual real wealth differ. Such wealth adjustment saving could be expressed as:

$$z = z\{w^d[(y_d^a)^*] - w\}; \qquad z_1 > 0,\ z(0) = 0. \tag{8}$$

Under such behavior, differences between desired and actual wealth would cause saving or dis-saving which causes desired and actual wealth to converge (by output or price level changes) until the stationary state is restored. Such saving behavior does not accord well with that assumed in the usual IS-LM analysis; but from equation (7) it is clear that growth can provide a continuing motive for saving in a "moving" equilibrium. Differentiating (7) with respect to time, we obtain:

$$\frac{dw^d}{d\tau} = w_1 \frac{d(y_d^a)^*}{d\tau}. \tag{9}$$

Continuing growth in permanent income and therefore desired real wealth can be justified for two reasons: first, because growth in per capita income will provide that subsequent generations will have higher permanent incomes than the generations they replace and, second, because the population of wealth holders is likely to be growing. Letting $(y_d^a)^*$ be determined by the modified adaptive expectations equation suggested by Friedman (1957, p. 144):

$$(y_d^a)^* = \beta \int_{-\infty}^{\tau} e^{(\beta-\gamma)(u-\tau)} y_d^a(u)\, du, \tag{10}$$

where γ is a constant secular growth term (assumed to be that of full employment output) and $\beta > \gamma > 0$ is a constant adjustment coefficient, we obtain:

$$\frac{d(y_d^a)^*}{d\tau} = \beta[y_d^a - (y_d^a)^*] + \gamma(y_d^a)^*. \tag{11}$$

Equations (9) and (11) imply a continuing flow equilibrium savings motive depending on current disposable income (adjusted) where:

$$\frac{dw^d}{d\tau} = \psi[y_d^a : (y_d^a)^*]; \qquad \psi_1 > 0. \tag{12}$$

By combining the desired flow saving in (12) with the stock adjustment saving in (8), we obtain a saving function of the form:

$$s^a = s^a[y_d^a, w : (y_d^a)^*]; \qquad s_1^a > 0,\ s_2^a < 0. \tag{13}$$

This saving function represents the flow demand for new wealth and therefore includes all changes in perceived net private wealth. The level of real wealth perceived by the private sector enters as a saving argument because of the stock adjustment motive, but this wealth dependency represents a temporary adjustment motive and will be ignored for simplicity. Note also that permanent income is now entered explicitly as a parameter giving rise to the short-run savings function (which assumes $[y_d^a]^*$ constant) and the long-run saving function (where $y_d^a = [y_d^a]^*$).

The flow equilibrium conditions are completed by introducing the flow supply of real wealth. Since I will be concerned only with macroeconomic equilibria where commodity and asset prices are unchanging over time, it is necessary only to consider the flow supply of new wealth at constant prices and interest rates. We can then differentiate $w = (1/P)(M^s + kB + E)$ with respect to time to obtain:

$$\frac{dw}{d\tau} = \frac{1}{P}\left(\frac{dM^s}{d\tau} + k\frac{dB}{d\tau} + \frac{dE}{d\tau}\right). \tag{14}$$

The flow supply of government bonds equals the government deficit minus that part of it financed by printing money plus the value of bonds exchanged in open market operations. Hence,

$$\frac{dB}{d\tau} = G - T' - \frac{dM_g}{d\tau} - \frac{dM_b}{d\tau} = G - T' - \frac{dM}{d\tau} \tag{15}$$

where $G = P \cdot g$ and $T' = P \cdot t'$ denote nominal values of government expenditures and taxes net of transfers. The flow supply of new equities equals the value of investment expenditures or:

$$\frac{dE}{d\tau} = P \cdot i(r : K^d) \qquad i_1 < 0 \tag{16}$$

where the marginal efficiency of investment function now explicitly includes the desired physical capital stock K^d as a parameter. As labor force growth and technological advances raise K^d, the marginal efficiency of investment curve will shift to the right in the i, r plane over time.

Substituting equations (15) and (16) into (14), we can equate the actual flow supply of private sector wealth to the flow demand as given by (13) to obtain:

$$s^a[y_d^a:(y_d^a)^*] = i(r:K^d) + k(g - t') + (1 - k)\frac{1}{P}\frac{dM}{d\tau}. \tag{17}$$

Equation (17) represents the flow equilibrium conditions and is easily seen to be equivalent to equation (6) except that certain parameters in the saving and investment functions are now explicitly listed.

III. Monetary and Fiscal Policy in Two Special Cases

When the future tax liabilities required to service the government debt are totally ignored, $k = 1$ and government bonds are fully perceived as collective net wealth by the private sector. In this case, equation (17) (or [6]) reduces to the standard IS curve expression given by (4) since $y_d^a = y_d$ by equation (5′). Thus, the standard IS-LM analysis can be derived from a more general model when government bonds are fully perceived as collective net wealth. In this case, a "pure" monetary policy which alters the money stock through open market operations involves only the exchange of one form of wealth (money) for another (bonds), consequently, the flow supply of wealth is not directly affected and monetary policy is constrained to altering the equilibrium level of income by shifting the LM curve due to its effects on the portfolio balance between monetary and nonmonetary assets.[11] When monetary policy must work through these channels, the possibility of weak monetary policy arises if private expenditure is insensitive to interest rate changes or when liquidity trap conditions prevail.

Fiscal policy, however, is fully effective when $k = 1$ because a change in the size of the government deficit alters the flow supply of wealth and shifts the IS curve horizontally by a multiplied amount. The ultimate change in y resulting from fiscal policy changes will be limited by the positive slope of the LM curve in the usual way; hence the expenditure theorists' emphasis on this parameter as a determinant of the effectiveness of fiscal policy.

When government bonds are not collective wealth of the private sector, the results are dramatically changed. If the tax liabilities attached to new bond issues are fully capitalized, then $k = 0$ and equation (17) becomes:

$$s^a\left[y - g + \frac{1}{P}\frac{dM}{d\tau}:(y_d^a)^*\right] = i(r:K^d) + \frac{1}{P}\frac{dM}{d\tau}. \tag{17′}$$

---

[11] In fact, as Metzler (1951) noted, there can be wealth effects of such monetary policy even if government bonds are treated as wealth. Expansionary monetary policy reduces interest rates and therefore raises the prices of nonmonetary assets increasing actual wealth in eq. (3). When the wealth argument is included in the savings function, this wealth effect increases its interest elasticity.

It is immediately apparent that the flow supply of new money balances enters the flow equilibrium condition regardless of how this flow supply enters the economy. Monetary policy will now alter the flow equilibrium conditions for two reasons. To begin with, the once and for all change in the stock of nominal money balances will alter private sector wealth since money, which is treated as net private wealth will be exchanged for bonds which are not. This gives rise to a temporary stock adjustment effect which is ignored above. More important, there will be a change in the flow supply of money balances since $dM/d\tau = \gamma M(t)$ where γ is the rate of monetary expansion, which is assumed constant and equal to the exogenous growth rate of full employment income. Consequently, monetary policy will now shift the IS curve even if the increased flow supply of money is engineered by open market operations. In particular, expansionary open-market operations will increase the flow supply of new wealth because the increased flow supply of money balances will be reducing the interest bearing government debt (and the tax liabilities needed to service that debt) at a faster rate. Since such monetary policy now involves a shift in both the IS and LM curves, its effectiveness does not depend critically on the interest elasticity of those curves.

When $k = 0$, fiscal policy is severely weakened since government expenditure financed by bonds is exactly equivalent to that financed by taxation. For this reason changes in tax rates have no effect whatever on the IS curve unless they result in monetary changes. Changes in g will still shift the IS curve but only by the change in g itself and not with a multiplier effect; therefore, fiscal policy is still weakly effective because of the balanced budget multiplier theorem which obtains because the marginal propensity to spend on domestic goods is unity for the government while income is partly saved in the private sector. It might be noted that this shift in the IS curve may be even smaller if government expenditures fall partly on imports and/or new public goods provide direct substitutes to private consumption causing "first-level" crowding out. In any event, the ultimate change in equilibrium income will always be smaller than the change in g because the LM curve is positively sloped.

Finally, one might expect the intermediate case where $0 < k < 1$ to be most likely. In this case fiscal policy is less effective than implied by the standard IS-LM model while monetary policy, although less potent than a dyed-in-the-wool monetarist might suspect, is nevertheless more effective than suggested by standard IS-LM analysis and does not strictly depend on the interest elasticities of the IS and LM curves as income expenditure theorists assert. In a slightly different vein, the results might be intermediate in the sense that a short-run versus long-run distinction may be important. In the short run it is probably reasonable to assume myopia on the part of the private sector with regard to the tax liabilities on changes in the government debt. Hence $k = 1$ and standard IS-LM results obtain. In the long run, however, these tax payments are likely to be fully perceived, so $k = 0$ will be the most appropriate assumption. This suggests that the standard IS-LM results are temporary and eventually yield to monetarist conclusions in the long run.

IV. IS-LM Analysis as Stock-Flow Equilibrium in a Growing Economy

In order to abstract from growth considerations, there is an understandable but unfortunate tendency to interpret the IS-LM equilibrium as a stationary one.[12] This has posed several difficulties which have perhaps impeded rather than aided the understanding of the equilibrium described. To begin with, it is difficult to reconcile the saving and investment behavior implied by the IS equation with a constant level of income except during a temporary stock adjustment phase. Moreover, it has led to the general presumption that a neutral monetary policy is one which maintains a constant money stock. In addition to introducing a misleading asymmetry to monetary and fiscal policy (because the former entails once and for all changes in the constant stock of money balances while the latter entails changes in the flow supply of government debt) this presumption has also contributed to a neglect of the flow supply of new money in the flow equilibrium conditions. This paper has argued that this omission is only tenable on the assumption that government bonds are completely perceived as net collective wealth. The interpretation of the IS-LM equilibrium in this paper involves growth as a basic element underlying the saving and investment decisions of the private sector, and this growth context also provides that neutral monetary and fiscal policy are consistent with nonzero flow supplies of monetary assets and government bonds.[13] The purpose of this brief section is to clarify the meaning of macroeconomic equilibrium and the nature of monetary and fiscal policy in a growing economy.

For a given price level and using equation (5), equations (1) and (17) can be plotted as IS and LM curves in the r, y plane with their intersection determining the current values of y and r assuming that equilibrating forces act instantaneously. In addition to the policy parameters $g - t'$, dM^s/dt and M^s, the positions of these curves depend on the exogenous parameters $(y_d^a)^*$ and K^d in the saving and investment functions. In addition to the equilibrium output level y, the existing stocks of productive inputs in this economy determine a full employment output level y_F which can be compared to the equilibrium level to obtain the factor utilization (or unemployment) rate. While it is usual to treat these values as constant over time, it is reasonable to assume that y_F is actually increasing over time as the labor force, capital stock, and level of technology increase. I will assume that y_F expands at the exogenous rate γ over time; thus, to prevent a continuing decline in the factor utilization rate, equilibrium y should also be increasing at the same rate.

---

[12] There are numerous exceptions, notably Mundell (1971, pp. 23-31) and others who have explicitly introduced growth considerations into the IS-LM framework.

[13] This is in direct contrast to the static approaches which have explicitly introduced government budget constraints to prevent the long-run flow supply of money (Christ [1968]) and bonds (Blinder and Solow [1973]) from being nonzero.

The above leads to a concept of growth equilibrium in which the IS and LM curves shift secularly over time as to increase y at the exogenous rate γ. The conditions for such a growth equilibrium at a constant interest rate can be found by taking the log time derivatives of equations (10) and (17) and substituting to obtain:

$$\varepsilon_{sy} Dy_d^a + \varepsilon_{sy^*}\left[\beta\left(\frac{y_d^a}{(y_d^a)^*} - 1\right) + \gamma\right] = \alpha_1 \varepsilon_{iK} DK^d + \alpha_2 D(g - t') + \alpha_3 D\left(\frac{dM^s}{d\tau}\right) \tag{18}$$

where D denotes the operator $dln/d\tau$, $\sum_{i=1}^{3} \alpha_i = 1$ and ε_{uv} is the partial elasticity between variables u and v. Neutral monetary and fiscal policy in a growing economy will be defined as a situation where $D(dM^s/d\tau)$ and $D(g - t')$ equal the full employment growth rate γ. Neutral monetary policy obtains when M^s grows at the constant rate γ since then $dM^s/d\tau = \gamma M^s$ Neutral fiscal policy obtains when government expenditures and taxes both grow at the full employment rate.

From equation (5′) it is clear that neutral monetary and fiscal policies imply that $Dy_d^a = Dy$. Furthermore, if the desired physical capital stock grows at γ and $\varepsilon_{iK} = 1$, it is clear that the right-hand side of (18) equals γ.[14] Finally it can be noted that Kuznets' study suggests that the long-run savings function is proportional, hence $\varepsilon_{sy} + \varepsilon_{sy^*} = 1$ and equation (18) implies that $Dy = \gamma$ when $y_d^a = (y_d^a)^*$. In this case, the IS curve moves to the right at the rate γ in the r, y plane and can be drawn as a stationary locus in the $r, y/y_F$ plane of figure 1.

The portfolio balance conditions can also be represented as an LM curve in figure 1, but a constant money supply would mean that the LM curve would be continually shifting to the left in that plane. The conditions for a stationary LM locus are found by differentiating (1) holding r constant to obtain:

$$Dy = \varepsilon_{l,y}^{-1} \qquad DM^s = \gamma/\varepsilon_{l,y}. \tag{19}$$

If the popular assumption that $\varepsilon_{l,y} = 1$ obtains, the LM curve will be stationary in the $r, y/y_F$ plane of figure 1 when monetary policy is neutral

---

[14] A further requirement in growth equilibrium is that the ratio of investment to the capital stock (I/K) equal γ. When $I/K \neq \gamma$, the capital stock will accumulate at a more or less rapid rate than competing inputs, hence diminishing returns will change K^d and shift the marginal efficiency of investment curve until $I/K = \gamma$ obtains.

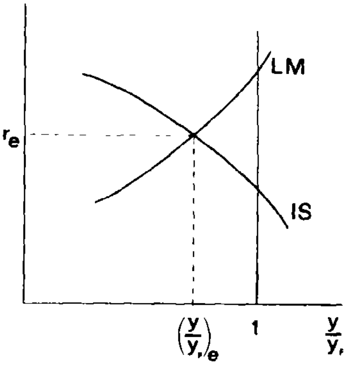


FIG. 1

as defined. If $\varepsilon_{l,y} \neq 1$, it is clear that a growth equilibrium with a constant interest rate is not possible with the IS equation of this model.

The intersection of the IS and LM curves determines the macroeconomic equilibrium at $(y/y_F)_e$ and r_e in figure 1, but this equilibrium now has an explicitly dynamic interpretation. Monetary policy involves changes in the stock of money balances (but not its rate of growth),[15] and this will increase the flow supply of monetary assets, since $dM^s/d\tau = \gamma M^s$. This monetary policy will shift the LM curve in figure 1 and possibly the IS curve, depending on whether k is less than or equal to unity. Changes in fiscal policy involve changes in the level but not the rate of growth of $g - t'$. This will shift the IS curve by an amount which depends on the value of k and on whether g or t' is altered.

V. Conclusions

The main argument of this paper is that saving behavior should be viewed as the desired accumulation of wealth by the private sector. As a result, the flow equilibrium conditions of the standard IS-LM model should be broadened to include certain net wealth changes which are ignored in the current formulation. This is particularly true if the private sector capitalizes some part of the tax liabilities attached to government bonds since open market operations will now alter the flow equilibrum conditions as well as the portfolio balance equation. As a result, it is found that the conclusions regarding the relative effectiveness of monetary and fiscal policy depend largely on whether or not government bonds are collective wealth to the private sector. If government bonds are not net wealth (the "rationalist" view) then monetarist conclusions regarding stabilization policies follow from the modified IS-LM model, while the standard IS-LM conclusions can be associated with the assumption that bonds are net wealth (the myopic view).[16] In addition it is hoped that by sketching out an interpretation of the IS-LM model as a dynamic wealth choice theory as done in this paper the focus of the monetarist debate can be sharpened and the differences between the two sides clarified.

---

[15] In the real world, these changes in the money stock will involve *temporary* rates of growth of the money supply different than γ. It should be noted that permanent rates of growth of M^s different from γ are inconsistent with equilibrium in this constant price level economy. Such monetary changes would shift the curves continuously until price level changes are induced and eventually some inflationary or deflationary equilibrium is established.

References

Barro, R. J. "Are Government Bonds Net Wealth?" *J.P.E.* 82, no. 6 (November/December 1974): 1095–1117.

Blinder, A. S., and Solow, R. M. "Does Fiscal Policy Matter?" *J. Public Econ.* 2 (1973): 319–37

Buchanan, J. M. "Barro on the Ricardian Equivalence Theorem." *J.P.E.* 84, no. 2 (April 1976): 337–42.

Christ, C. "A Simple Macroeconomic Model with a Government Budget Constraint." *J.P.E.* 76, no. 1 (January/February 1968): 53–67.

Drèze, J., and Modigliani, F. "Consumption Decisions under Uncertainty." *J. Econ. Theory* 5 (1972): 308–35.

Feldstein, M. "Perceived Wealth in Bonds and Social Security: A Comment." *J.P.E.* 84, no. 2 (April 1976): 331–36.

Frenkel, J. A., and Rodriguez, C. A. "Portfolio Equilibrium and the Balance of Payments." *A.E.R.* 65, no. 4 (September 1975): 674–88.

Friedman, M. *A Theory of the Consumption Function.* Princeton, N.J.: Nat. Bur. Econ. Res., 1957.

———. *A Theoretical Framework for Monetary Analysis.* NBER Occasional Paper no. 12. New York: Columbia Univ. Press, 1971.

———. "Comments on the Critics." *J.P.E.* 80, no. 5 (September/October 1972): 906–50.

Metzler, L. "Wealth, Saving, and the Rate of Interest." *J.P.E.* 59, no. 2 (April 1951): 93–116.

Mundell, R. A. *Monetary Theory.* Pacific Palisades, Calif.: Goodyear, 1971.

Patinkin, D. *Money, Interest and Prices.* 2d ed. New York: Harper & Row, 1965.

Purvis, D. "Portfolio and Consumption Decisions: Towards a Model of the Transmission Process." Unpublished manuscript, September 1975.

Shell, K.; Sidrauski, M.; and Stiglitz, J. E. "Capital Gains, Income and Saving." *Rev. Econ. Studies* 36 (January 1969): 15–26.

Silber, W. L. "Fiscal Policy in IS-LM Analysis: A Correction." *J. Money, Credit and Banking* 2 (November 1970): 461–72.

Tobin, J. "A General Equilibrium Approach to Monetary Theory." *J. Money, Credit and Banking* 1 (February 1969): 15–29.

———. "Friedman's Theoretical Framework." *J.P.E.* 80, no. 5 (September/October 1972): 852–63.

Witte, J. G. "The Microfoundations of the Social Investment Function." *J.P.E.* 71, no. 5 (October 1963): 441–56.

---

[16] In this light, the recent attempts to analyze this question more formally (Barro 1974, Feldstein 1976, and Buchanan 1976) are most welcome.

Price Discrimination and Vertical Control: A Note

J. R. Gould

London School of Economics and Political Science

The analysis extends beyond the standard case of the monopolist who integrates in order to isolate some of his customers from arbitrage in the monopolized good so that he can discriminate against them. It examines the exploitation of factors employed by, and the customers of, the monopolist's customers. By vertical integration the monopolist may utilize barriers to arbitrage that would otherwise be unexploited because the market is competitive. Sometimes looser forms of vertical control, for example, tying arrangements, can be employed, and in some circumstances it is possible to extract surpluses from parties with whom the monopolist does not trade.

Although it is widely recognized that a monopolist might integrate forward to gain the benefits of price discrimination, discussion of the proposition appears to go little further than explaining that vertical integration can be used to isolate segments of the market in the absence of other barriers to arbitrage among the monopolist's customers. This note attempts a more systematic analysis; it extends beyond discrimination among the monopolist's own customers to examine (i) the exploitation of (*a*) the factors employed by his customers and (*b*) the customers of his customers, and (ii) whether vertical integration is necessary to exploit parties with whom the monopolist does not trade.[1] When arbitrage in a

---

[1] The stimulus for these extensions comes from a somewhat Delphic aside in Stigler (1951, p. 191): "A firm cannot practice price discrimination in the stages in which it does not operate; only by fabricating cable could the Aluminum Company of America sell cable at less than the ingot price, while maintaining a higher price on less competitive products. . . . " The connection between the general proposition and the illustrative case is not clear: the point of the latter would seem to be that ALCOA integrated because it could not discriminate in the market in which it *did* operate, the ingot market, the prevention of resale of ingots being impossible. Of more importance, it will be argued—contrary to the general proposition—that the gains from discrimination in remoter markets can in some circumstances be achieved without recourse to vertical integration. However, the analysis lends support to Stigler's remark, preceding the above quotation, that "Monopoly is a devious thing. . . . "

[*Journal of Political Economy*, 1977, vol. 85, no. 5]
© 1977 by The University of Chicago. All rights reserved.

good or service is, or can be made, impossible or prohibitively expensive, there is, so to speak, a potential for discrimination that cannot be used by traders in the market if it is perfectly competitive. This potential may in some circumstances be exploited by a firm with monopoly power at a related stage.

The Model

Apart from discrimination, the principal explanations that have been advanced for vertical integration by a monopolist are the following: to avoid the cumulative restrictions of output of successive monopoly and the bargaining costs of bilateral monopoly (e.g., Machlup and Taber 1960); to avoid distorting factor proportions in the customer industry (e.g., Vernon and Graham 1971; Warren-Boulton 1974); or if there are "cost advantages," that is, if the coordination of stages of production by administration is efficient relative to market transactions (e.g., Coase 1937; Williamson 1971).

If these other reasons for vertical integration are ruled out by assumption, the relevance of discrimination can be brought into sharper focus. Accordingly, we assume initially that a monopolist producer of input A sells only to competitive industry X; that production of X is subject to fixed coefficients, with units defined so that one unit of X requires inputs of one unit of A and one unit of B; that input B is competitively supplied; and that no cost advantages or disadvantages would result from the integration of the production of X and A.

This model has been used (by, e.g., Needham 1969, pp. 118–19; Singer 1968, p. 208) to argue that abstracting from price discrimination the A monopolist has no incentive to integrate into X production. The analysis is outlined with reference to figure 1. The monopolist's derived demand curve D_a is obtained by subtracting vertically the supply curve of input B from the demand curve for X. Thus $D_a = D_x - S_b$; and it follows that $MR_a = MR_x - MS_b$, where MS_b is the curve marginal to S_b. For maximum profits

$$MC_a = MR_a = MR_x - MS_b. \tag{1}$$

The monopolist's profit-maximizing price is P_a^*. The competitive supply curve of X is $S_b + P_a^*$, yielding an equilibrium price of P_x^*. The net profit to the system, all accruing to the A monopolist, is the revenue $P_x^* X_1^*$ minus the cost of producing A_1^* and of acquiring B_1^*.

Now suppose that the A monopolist integrates and monopolizes X production, cost and demand conditions unchanged. The integrated firm's marginal cost curve is $MC_a + MS_b$, and the condition for maximum profits,

$$MC_a + MS_b = MR_x, \tag{2}$$

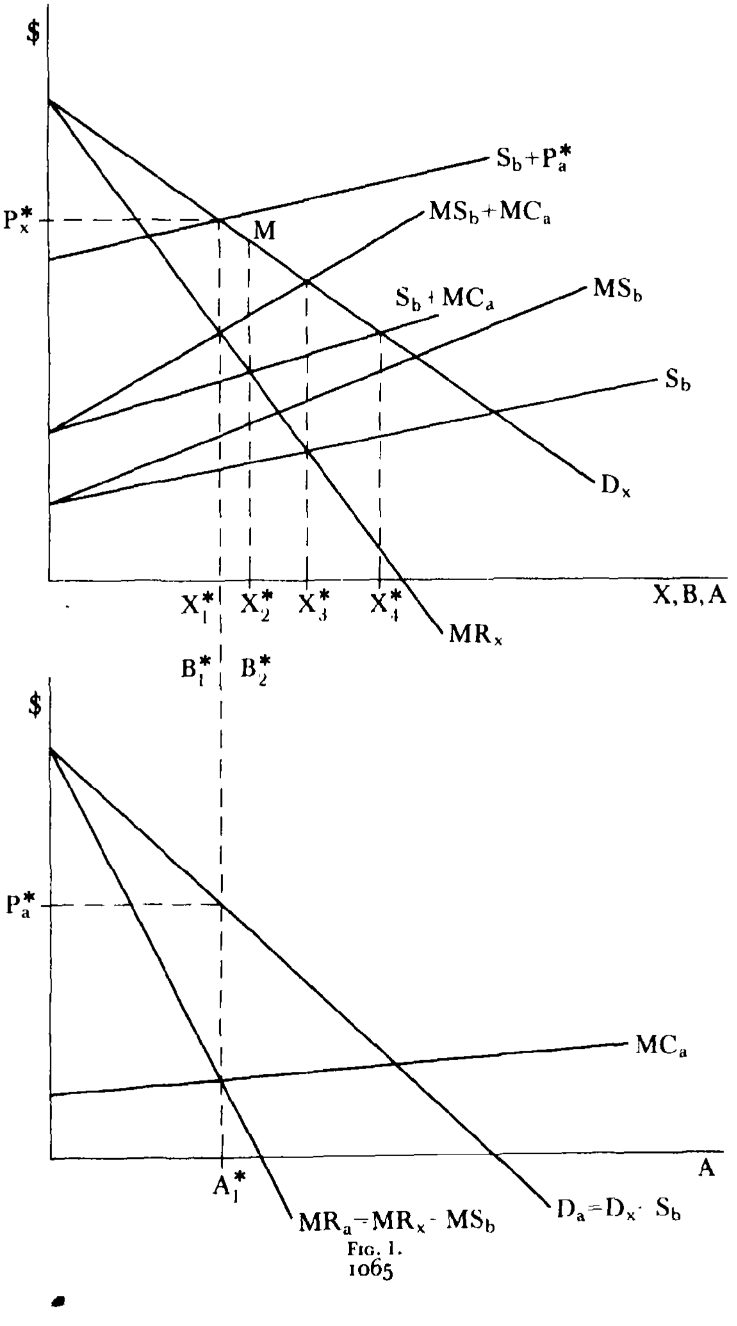

$
$S_b+P_a^*$
P_x^*
M
MS_b+MC_a
S_b+MC_a
MS_b
S_b
D_x
X_1^*
X_2^*
X_3^*
X_4^*
X, B, A
MR_x
B_1^*
B_2^*
$
P_a^*
MC_a
A
A_1^*
$MR_a=MR_x-MS_b$
$D_a=D_x-S_b$

FIG. 1.

is the same as condition (1). Thus output and price of X, and the aggregate cost and revenue to the system, are the same as in the nonintegrated case—there is zero net gain from vertical integration.

Notice that conditions (1) and (2) indicate that the nonintegrated monopolist takes full account of the slopes of both D_x and S_b. In other words, he exercises—indirectly via the price of A—monopoly power over the buyers of X and monopsony power over the sellers of B just as effectively as if he had been fully integrated into X production and traded directly in the markets for X and B. On the other hand, although the monopoly power over A is exerted in the price charged to the X firms, the X entrepreneurs, in this model, earn zero profits in the long run, just as they would if the input A were supplied to them by a competitive industry. To borrow terminology from the theory of taxation, the impact of the monopolist's power is on his customers, but the incidence is on the customers of, and the factors employed by, his customers.

Discrimination among Customers

Discussions of the relation between vertical integration and price discrimination argue informally that vertical integration is a surrogate for discrimination when arbitrage among customers cannot be prevented (e.g., Stigler 1951, p. 191; Needham 1969, p. 116). It will be useful to extend the fixed-proportions model to a case of discrimination corresponding to that in the literature. The additional point to be established is that discrimination by the nonintegrated monopolist is analytically equivalent to the exercise of *simple* monopoly-monopsony power by the fully integrated monopolist.

Assume that the nonintegrated A monopolist sells to the X industry, described above, and to the competitive Y industry; that production of Y *requires fixed proportions of inputs A and C; and that the demands for* X and Y are unrelated. The Y industry's demand for input A is $D_y - S_c$, with marginal revenue $MR_y - MS_c$. Given that the A monopolist can discriminate between the X and Y industries, his profit-maximizing conditions are

$$MC_a = MR_x - MS_b = MR_y - MS_c. \qquad (3)$$

These conditions determine the profit-maximizing output of A and its division between the X and Y industries, the prices charged being then determined from the derived demand curves.

If the A monopolist cannot discriminate because of absence of barriers to resale, an equally profitable result can be achieved by integrating and monopolizing the production of the good that would have enjoyed the lower price for input A under discrimination, say X, and selling A to independent Y producers. The profit-maximizing conditions are then

$$MC_a + MS_b = MR_x, \qquad \text{and } MC_a = MR_y - MS_c, \qquad (4)$$

equivalent to (3).

Alternatively, the A monopolist could integrate and monopolize the production of both X and Y. The profit-maximizing conditions then become

$$MC_a + MS_b = MR_x, \qquad \text{and } MC_a + MS_c = MR_y, \qquad (5)$$

which are the same as (3) and (4).

The equivalence of conditions (3), (4), and (5) indicates that each case results in the same quantities of inputs and outputs, and the same net revenue to the system, always accruing to the A monopolist. From another point of view, the equivalence of conditions (3) and (5) shows that if the nonintegrated monopolist *can* discriminate among buyers of A the results are the same as if he had been fully integrated into the production of both X and Y and exercised *simple* monopoly power over his customers and *simple* monopsony power over the sellers of the factor services B and C. There is nominal discrimination when the nonintegrated monopolist charges a higher price to the Y firms than to the X firms, but the X and Y entrepreneurs earn zero profits in the long run, in this model, whether or not the A monopolist discriminates. The effects of the nominal-discrimination are borne by parties with whom the monopolist does not trade.

Discrimination among Factors Employed by Customers

We have seen that the nonintegrated monopolist can exercise—indirectly through the price, or discriminatory prices, he charges for A—simple monopsony power over other factors employed by his customers. This observation prompts the question: how and in what circumstances can the A monopolist gain the additional benefits of discriminating monopsony in these markets?

For simplicity we revert to the assumption that the A monopolist sells only to the competitive X industry, which uses input B. Assume that there are barriers to arbitrage among the sellers of B. The potential for discriminatory exploitation depends on whether the sellers of B earn rents. Consider two possible reasons for the upward slope of S_b. First assume that all units of B are equally efficient, that B is used by several industries, and that the X industry accounts for a sufficiently large proportion of the market for the price of B to be bid up as the X industry expands. Discrimination among units of B is impossible on these assumptions. An integrated monopolist would have to pay to each unit of B the price necessary to bid the marginal unit away from other industries. The integrated monopolist is limited to the exercise of simple monopsony power over B; but, as we have seen, a nonintegrated monopolist can exploit this power just as effectively by choosing the appropriate price for A. The assumed supply conditions for B provide no incentive to integrate vertically.

Alternatively, suppose that units of B, although equally efficient in the production of X, are of differing efficiencies in other industries. Efficiency differentials will be established among the prices of units of B employed outside the X industry. The units of B employed by the X industry will be those that are less efficient in other industries. The X industry, as it expands, will have to pay successively higher prices for marginal units of B. But since units of B are equally efficient in the production of X, competitive bidding by independent X firms will establish a common price equal to the transfer earnings of the marginal unit of B, the intramarginal units receiving a surplus over transfer earnings. On the other hand, a vertically integrated monopolist of A and X exercising perfect discrimination would pay to each unit of B only its transfer earnings. His marginal cost of B is then represented by S_b in figure 1, and his equilibrium output is X_2^*, where $MR_x = S_b + MC_a$. (Evidently, similar possibilities for exploitation could arise in a model with units of B unequally efficient in the X industry.)

However, vertical integration into the production of X is not essential for the exploitation of the B factor rents. The A monopolist can—on present assumptions establish himself as the monopolist seller of B to the X industry by stipulating that the X firms buy all their requirements of B from him, that is, by tying sales of A to sales of B. This would make him the monopsonist buyer of B for employment in the X industry. He could then offer the units of B that are less efficient in other industries just sufficient to induce them to work in the X industry, while charging the X firms a common price for B equal to the transfer earnings of the marginal unit at the quantity B_2^*.[2]

A nonintegrated monopolist of A need not even establish direct contractual relations with the sellers of B in order to extract their rents if he can discriminate between the X firms. The objective can be achieved by suitable pricing of his own good A. Taking the vertical heights of S_b in figure 1 to represent the transfer earnings of successive units of B, the monopolist charges the X firms for each unit of A the difference between X_2^*M and the transfer earnings of the unit of B with which that unit of A

---

[2] Compare Bowman (1957) and Burstein (1960) on tie-in sales. These analyses relate to the extraction of the surplus on the tying good, whereas in our case the surplus on the tied good is extracted. This distinction is crucial in the context of Bowman's discussion of the concept of leverage in antitrust law. Also, Bowman's discussion of tying as a device to measure intensity of use for purposes of price discrimination leads him to the conclusion that there is no advantage in tying when the goods are used in fixed proportions.

cooperates. The *X* firms could then bid no more for each unit of *B* than its transfer earnings; the nominal discrimination is in the prices of *A* to the *X* firms, but the burden falls on the sellers of *B*. Such indirect exploitation depends on information about the *B* input that would often be difficult for the nonintegrated monopolist to acquire, but in some circumstances, for example, if the *B* inputs were land of different fertilities, estimation of the surpluses of inputs and identification of their employers might well be feasible.

A further point worth noting is that in this last case the *X* firms have identical demand curves for *A* and earn no surpluses. Thus the observation that a firm is discriminating among its customers does not necessarily entail the inference that it is exploiting either differences in their demand elasticities or their surpluses.

The analysis points to several conclusions. If there are barriers to arbitrage in a factor market, or if they can easily be erected, the existence of rents to factors employed by a competitive customer industry provides an incentive for a monopolist to integrate vertically. However, vertical integration may not be necessary to exploit factor rents if some looser form of vertical control can be operated, such as a tying arrangement or discrimination among the monopolist's customers. Indeed, a monopolist may be able to secure the gains of discrimination without contractual relations with the parties being exploited. The choice of method of exploitation will be influenced by considerations other than those mentioned above and from which the model abstracts, such as the costs of making and ease of enforcing the alternative contractual arrangements, and the information about factor rents that they generate. Perhaps the most likely potential for exploitation lies in differential site rents. For example, the exploitation of owners of retail sites with special advantages for a particular trade may provide a possible rationalization of some cases of vertical integration into retailing, and of the practice of leasing retail premises to independent operators, subject to tying clauses; this is an alternative hypothesis to the more usual explanations in terms of cost savings, or of barriers to entry into the monopolist's stage.

Discrimination among Customers of Customers

This case warrants only brief discussion, for the analysis and conclusions are analogous to those of the previous case. If the buyers of *X* are firms, then the monopolist of *A* may integrate into production of *X* in order to exploit either the firms themselves, or the factors they employ, or their customers. For simplicity, the following comments assume that the buyers of *X* are households.

Suppose that *X* is a good or service that it is impossible or prohibitively costly for buyers to resell to each other. Competitive producers of *X* cannot make use of these barriers to resale, for discrimination is inconsistent with perfect competition. However, the monopolist of A can, by integrating and monopolizing the production of X, combine the monopoly power and barriers to retrading necessary to exploit the potential for discrimination. Taking the limiting case of perfect discrimination—by, say, a two-part tariff—and assuming negligible income effects, the integrated monopolist's marginal revenue curve is D_x. Assuming that it is impossible to discriminate between sellers of B, the integrated marginal cost curve is $MS_b + MC_a$, and the profit-maximizing output is X_3^*. If it is possible to discriminate perfectly against sellers of B, the profit-maximizing output is X_4^*.

Integration into the production of X may not be necessary to exploit the buyers of X. As a possible alternative the monopolist could make sales of A to the X firms conditional on the X firms selling all their output to him He thus would become the sole seller to buyers of X and could discriminate among them. Such a contract is similar to one for the processing of A and is, of course, by no means unusual.

Finally, it may be possible for the A monopolist to discriminate indirectly among buyers of X, without trading with them, if he can discriminate among the X producers. For example, if the buyers of X are spatially separable, the A monopolist could make sales of A conditional on each X firm selling only in a particular area, that is, operate the demands for X by discriminatory pricing of A to the X firms. In effect, he creates local monopolies to enable the X firms to exploit their customers, and then captures the spoils from the X firms.

This paper has been concerned with the relation between discrimination and forward vertical control, but it is evident that potentialities for discrimination give rise to similar incentives to backward integration and to possibilities of indirect exploitation.

References

Bowman, W. S., Jr. "Tying Arrangements and the Leverage Problem." *Yale Law J.* 67 (November 1957): 19–36.

Burstein, M. L. "The Economics of Tie-In Sales." *Rev. Econ. and Statis.* 42 (February 1960): 68–73.

Coase, R. H. "The Nature of the Firm." *Economica*, n.s. 4 (November 1937): 386–405.

Machlup, F., and Taber, M. "Bilateral Monopoly, Successive Monopoly, and Vertical Integration." *Economica*, n.s. 27 (May 1960): 101–19.

Needham, D. *Economic Analysis and Industrial Structure.* New York: Holt, Rinehart & Winston, 1969.

Singer, E. M. *Antitrust Economics: Selected Economic Cases and Economic Models.* Englewood Cliffs, N.J.: Prentice-Hall, 1968.

Stigler, G. L. "The Division of Labor is Limited by the Extent of the Market." *J.P.E.* 59, no. 3 (June 1951): 185–93.

Vernon, J. M., and Graham, D. A. "Profitability of Monopolization by Vertical Integration." *J.P.E.* 79, no. 4 (July/August 1971): 924–25.

Warren-Boulton, F. W. "Vertical Control with Variable Proportions." *J.P.E.* 82, no. 4 (July/August 1974): 783–802.

Williamson, O. E. "The Vertical Integration of Production: Market Failure Considerations." *A.E.R.* 61 (May 1971): 112–23.

Comments

Pollak and Wachter on the Household Production Function Approach

William A. Barnett

Board of Governors of the Federal Reserve System

In "The Relevance of the Household Production Function and Its Implications for the Allocation of Time," Pollak and Wachter (1975) have provided a valuable analytical interpretation of the theory underlying the "new home economics." In addition, they have provided insights into serious potential abuses of the household production function approach. However, their identification of one such potential abuse led them to terminate their analysis prematurely with the rejection of the entire shadow-price concept on which much of the new home economics is based. Their conclusion is unwarranted.

Pollak and Wachter maintain that joint production is inherently important in household technology, and they argue that joint production breaks the link between the existing household production function approach and the neoclassical theory on which that approach is based. They also argue that joint production results in the confounding of tastes and technology within shadow prices. But such "confounding" could pose a fundamental theoretical problem only if the postulated "confounding" can be translated into an identification problem. I shall equate a particular theoretical structural model with the household production function approach. I shall demonstrate that all functions in that structural form do have known neoclassical properties, and I shall discuss the identification of the structure. I shall derive household structure

---

I have benefited from the comments of Richard Berner, John Davison, Kenneth Kopecky, and an anonymous referee. I have also benefited from discussions on the subject with Daniel Brod, Robert Pollak, and James Heckman. The views expressed herein are solely those of the author and do not necessarily represent the views of the Board of Governors of the Federal Reserve System.

[*Journal of Political Economy*, 1977, vol 85, no 5]
© 1977 by The University of Chicago All rights reserved.

and prove its identification when tastes are Bergson and technology is Hybrid Diewert. The structure will be shown to be overidentified, and joint production will be shown to increase the number of overidentifying restrictions. I argue that in the general case joint production commonly tends to assist in identification without introducing any nonneoclassical theoretical complications.

Having rejected commodity shadow prices, Pollak and Wachter recommend an alternative. I shall equate that alternative with a reduced-form approach not having capabilities comparable to those of the household production function approach.

I. Introduction

Following Pollak and Wachter, let $X = (x_1, \ldots, x_n)'$ be a vector of goods, and let $Z = (z_1, \ldots, z_m)'$ be a vector of "commodities" generated from goods by the household's production process. Let U be the household's utility function, which we shall assume is defined over commodity vectors, and let $P = (p_1, \ldots, p_n)'$ be the vector of goods prices. Pollak and Wachter have shown that a cost function $C(P, Z)$ exists such that the household maximizes $U(Z)$ subject to the constraint $C(P, Z) = \mu$, where μ is total expenditure available.[1] The solution is a system of commodity demand equations $Z = f(P, \mu)$. Translating U back into the goods space, we can also derive goods demand functions $X = h(P, \mu)$.

Assuming that household production is characterized by constant returns to scale, Pollak and Wachter have shown that the household can be shown equivalently to solve for that value Z^* which will

$$\text{maximize } U(Z) \text{ subject to } \pi'Z = \mu, \tag{1}$$

where $\pi = (\pi_1, \ldots, \pi_m)'$ is the gradient of $C(P, Z)$ with respect to Z and π therefore is a function of (P, Z). Then $\pi_i(P, Z)$ is defined to be the shadow price of the ith commodity.

As Pollak and Wachter have observed, much of the appeal of the commodity shadow-price approach lies in its ability to use functions having known neoclassical properties. However, Pollak and Wachter maintain that, if the constraint $\pi(P, Z)'Z = \mu$ is nonlinear in Z, then the link with conventional theory is broken, since commodity demand functions derived from (1) would "correspond to those in a model in which consumers are monopsonists or are offered tie-in sales" (p. 258). But the commodity shadow prices $\pi(P, Z)$ do depend upon Z whenever household production exhibits jointness, which Pollak and Wachter maintain is inherently characteristic of household production processes. Hence

---

[1] Pollack and Wachter have considered the additional inclusion of time in the model. The theoretical issues I shall discuss will be presented in such a manner as to be unchanged by such complications.

Pollak and Wachter conclude that in the usual case the household production function approach must model a nonneoclassical decision problem for which "there are virtually no substantive results" (p. 258).[2] On these grounds, which I shall show to be specious, they immediately reject the use of commodity shadow prices as arguments of commodity demand functions.[3] When Z is not measurable, they recommend the estimation of h. Otherwise they recommend estimation of f (perhaps preceded by a prior-stage estimation of technology).

Pollak and Wachter have observed correctly that their function f does not have the known properties of conventional neoclassical demand functions. In addition, f and h depend upon both technology and preferences in a manner that provides little information about either. But as Pollak and Wachter have explained, the primary objective of the new home economics is to avoid "confounding tastes and technology" (p. 260). I shall show that the use of commodity shadow prices permits us to isolate sources of taste and technological change while using *only* functions having known conventional neoclassical properties.

II. Basic Constructs

Shadow prices are usually defined in terms of the normal to a separating hyperplane constructed at a solution point. That construction is dependent upon the location of the solution point, which need not be solely supply or technology determined. Now recall that $Z^* = f(P, \mu)$ is the household's solution value for Z. Define π^* by $\pi^* = \pi(P, Z^*)$, and let us instruct the household to reselect Z conditionally upon π^* to[4]

$$\text{maximize } U(Z) \text{ subject to } \pi^{*\prime}Z = \mu. \qquad (2)$$

The constraint in problem (2) is the hyperplane contemplated by the shadow-price approach. The solution for Z (in terms of $[\pi^*, \mu]$) can be denoted implicitly by[5] $G(Z, \pi^*, \mu) = 0$. By comparing the first-order

---

[2] Observe carefully that the issue they raise is the availability of theoretical knowledge of the properties of the functions used in the household production function approach. This mathematical question about nonstochastic function properties is independent of the separate statistical question of the endogeneity of any random variables. E.g., the endogeneity of the random variable $\pi(P, Z)$ follows trivially from the direct functional dependency of $\pi(P, Z)$ upon the endogenous random variable Z. Yet this endogeneity is irrelevant to Pollak and Wachter's contention of the presumed unavailability of known theoretical properties of the function π or of any other function in household structure.

[3] In order to address these issues, I shall not exclude joint production from this paper.

[4] To permit our prior computation of π^*, I assume that the household already has solved its full decision problem (1). Nevertheless, problem (2) can be defined formally, despite the seeming redundancy of its objectives.

[5] Formally we could use the conventional notation $Z = g(\pi^*, \mu)$, but I use the implicit function notation G to emphasize the fact that an explicit closed-form solution for g need not always exist. Practical and relevant examples are provided in Barnett (1977; 1978, chap. 7)

Wachter have warned us against convincingly. But to use the shadow-price function, π, in the construction of ([3], [4]), we have no need to impute explanatory power to the value of the variables[12] $\pi(P, Z)$.

Each function in the reduced form can carry joint information both about preferences and technology. If we wish to investigate properties of the household's structure or to consider household structural change, we must use a structural parameterization permitting the unscrambling of tastes from technology.[13] To investigate technological change, we can explore shifts in the parameters of the function π and the function F. I do not explore variations in the value of $\pi(P, Z)$, since such variations depend upon preferences as well as technology. In brief, the merits of Pollak and Wachter's approach are precisely those of a reduced-form system, while those of the shadow-price approach are precisely those of a structural form. But the advantages of structural-form estimation are well known.[14]

When Z is measurable, Pollak and Wachter advocate estimating (5′) (perhaps preceded by the estimation of technology). The most general approach to modeling f would involve parameterizing f directly.[15] Each parameter of such a direct reduced-form parameterization could carry information about both preferences and technology, and untangling the two sources (whether or not technology is itself estimated in a prior stage) would rarely be feasible. This is truly a reduced-form approach. Alternatively, we could structurally parameterize (5′) by selecting parameterizations of preferences and technology (rather than directly of f) and then deriving (5′) in terms of those original structural parameters. But this approach would be of little practical value, since the resulting system would be derivable only in pathological cases,[16] and even in those rare cases the resulting system would be far more difficult to estimate than our model.[17]

---

[12] Different households may have identical technologies and be "given" identical (P, μ) without having identical shadow prices. In such a case, I should conclude that shadow prices are different as the result of differing tastes. To view shadow prices as explanatory would reverse the direction of causation.

[13] E.g., without a structural parameterization habit formation could not be incorporated into the model without confounding tastes and technology. Prior estimation of technology would not help. Of course in the exceptional borderline case of exact identification, structural-form parameters for a fixed structure can be computed from reduced-form parameters. But structural change can be investigated only in terms of changes in structural-form parameters. Furthermore, even if the structural form were exactly identified, nonlinearities that typically exist in both the structural and reduced form would severely complicate solution for the structural-form parameters from the reduced form.

[14] Furthermore, reduced-form forecasts are easily computed numerically from an identified structural form, so that the objectives of the reduced form can be served by the structural form itself. No need exists ever to estimate directly or to solve analytically for the inherently less informative reduced form. Also observe that the household's structural form ([3], [4]) is well designed for deriving refutable theoretical results. The properties of all the functions in ([3], [4]) are restricted by neoclassical demand and production theory, regardless of whether or not joint production exists. Those restrictions imply restrictions upon the response of Z to variations in tastes, technology, and (P, μ). But such theoretical results are very weak unless further assumptions are made about tastes and technology. In this context it is frequently productive to exclude corner solutions and inferior commodities. Excluding joint production is neither necessary nor desirable. In contrast, observe that Pollak and Wachter's function f itself does not possess conventional neoclassical demand properties, although its actual properties can be deduced from ([3], [4]).

[15] The selection of such a direct reduced-form parameterization preferably should be guided by a duality theory, although in practice we may be satisfied with a parameterization of f which only approximates the underlying theory.

IV. Identification of the Structural Form

The household production function approach can be viewed as predicated upon the ability of system ([3], [4]) to unscramble tastes from technology. The dependence of shadow prices both upon tastes and technology (which troubles Pollak and Wachter) could pose a fundamental methodological problem only if that joint dependency resulted in an identification problem; furthermore, my own advocacy of ([3], [4]) as the household's structural model would be unsupportable if it were unidentified. To dispel in advance any truly serious potential doubts about the use of shadow prices, I shall disprove nonidentification by counterexample.

Consider a two-good, two-commodity household. I shall assume that the household's commodity demand functions are of the Bergson form $z_i = \beta_i(\mu/\pi_i)$, with $\beta_i > 0$ for $i = 1, 2$. I shall assume that the household has a Hybrid-Diewert joint cost function.[18] Using a_{ijkl}, $i, j, k, l = 1, 2$, to denote parameters, we have

---

[16] To complete the system, suppose that we were to adjoin (4) to (5′). Then suppose that we were to attempt to derive the resulting system from a prior parameterization of tastes and technology. We could first derive ([3], [4]), which indexes the equivalence class of structural forms consisting of all elementary transformations of ([3], [4]). To pass from the shadow-price approach, defined (in the wide sense) by this equivalence class, to the Pollak and Wachter equations (5′), we must be able to solve the structural model (3) explicitly for a closed-form representation of the reduced form equations (5′). As is true in general for nonlinear structures, this rarely is possible. As a simple example, consider Hybrid-Diewert technology and CES commodity preferences in the two-good, two-commodity case, with the same notation introduced in Section IV below. Let $x = \sqrt{z_2}$ and let β be an arbitrary integer exceeding 4. Apply the binomial expansion to terms of order β, and collect all terms onto the right-hand side. To separate z_1 and z_2, we must be able to solve this polynomial for x. But the polynomial is full and of order $\beta + 1$, and it is well known in Galois theory that the general polynomial of degree exceeding 4 is not solvable (except in terms of Fuchsian functions, which are not empirically implementable). Hence we see that to parameterize ([4], [5′]) structurally we must back up to an implicit representation, i.e., into an element of the equivalence class defining the shadow-price approach itself.

[17] The sole such case of which I am aware (permitting both joint technology and a closed-form representation of [5′]) is the case of Cobb-Douglas preferences and Hybrid-Diewert technology. In that case, ([4], [5′]) is both nonlinear in the variables and deeply nonlinear in its parameters, while our system ([3], [4]) is nonlinear in the variables but fully linear in all of its parameters.

$$\begin{aligned} C(z_1, z_2, p_1, p_2) = {} & a_{1111} z_1 p_1 + a_{1122} z_2 p_1 + a_{2211} z_1 p_2 \\ & + a_{2222} z_2 p_2 + 2a_{1211} z_1 \sqrt{p_1 p_2} + 2a_{1222} z_2 \sqrt{p_1 p_2} \\ & + 2a_{1112} p_1 \sqrt{z_1 z_2} + 2a_{2212} p_2 \sqrt{z_1 z_2} \\ & + 4a_{1212} \sqrt{z_1 z_2} \sqrt{p_1 p_2}. \end{aligned}$$

We can then derive the household structural form[19] ([3], [4]). The rank (sufficient) condition for identification can be shown to be satisfied, and the system thereby is identified.[20] The order condition is satisfied for an equation if the number of exclusion restrictions on the equation exceeds the number of equations in the system by more than one. The number of exclusion restrictions on each equation is 14, and the number of equations in the system is four. The system is considerably overidentified.

I now shall explore the effect of nonjointness on identification. To impose nonjointness, we set $a_{1112} = a_{1212} = a_{2212} = 0$. Eleven exclusion restrictions remain in each of two equations of ([3], [4]), while 13 remain in each of the other two equations. We have lost three overidentifying restrictions in each of two equations and one overidentifying restriction in each of the other two equations. Joint production helps in identification.

The results acquired from my counterexample reflect general properties of the theoretical structural form ([3], [4]) rather than properties specific to the chosen specification. The large number of exclusion restrictions results from the nonlinearity in the variables inherent to ([3], [4]) and from the fundamental difference between the structures of[21] (3) and (4).

---

[18] See Hall (1973) for the definition of the Hybrid-Diewert cost function.

[19] Detailed verification of all of my results on this counterexample is contained in a prior draft of this paper, available upon request.

[20] Since the structure in this case is nonlinear only in its variables (rather than its parameters), the rank and order conditions for identification are available immediately from Fisher (1966, chap. 5)

[21] It is well known that nonlinear functions of exogenous variables can be viewed as new exogenous variables. Although the exogenous variables P may appear in all of the equations of ([3], [4]), some functions of the elements of P certainly will be missing from some equations. The result is exclusion restrictions on those equations whenever the functions do exist elsewhere in the system. Fisher (1966) has proved that terms involving endogenous variables can generate exclusion restrictions in a similar manner, and ([3], [4]) is inherently nonlinear in both its endogenous and exogenous variables. The fundamental difference between the structure of (3) and (4) assures the existence of many such exclusion restrictions in the combined system ([3], [4]). Furthermore, observe that μ occurs in (3) but not in (4). Each occurrence of μ alone or in an interaction with an endogenous or exogenous variable of (3) provides an exclusion restriction on (4). Finally, in my example I have ignored the 15 available cross-equation parameter restrictions. Such

Joint production does not hinder identification. In fact, it is known in general that such interactions and the nonlinearities which result do not hinder (and commonly assist in) identification.[22]

V. Conclusion

The household structural form that I have identified with the household production function approach does not and need not contain commodity shadow prices as predetermined or supply-determined variables to which causality can be imputed but contains them, rather, as functions of both the exogenous variables P and the endogenous variables Z. The household's structural form contains only functions having conventional neoclassical properties, with each function related solely and identifiably either to preferences or to technology. Causality can be imputed to explainable taste and technological change and to variations in the exogenous variables (P, μ). The existence of joint production poses no problems in the modeling of household structure.

Pollak and Wachter also discuss the formidable problems involved in defining and measuring commodity-consumption quantities. Those issues are independent of Pollak and Wachter's more fundamental critique of the theory underlying the household production function approach, and I have abstracted from such measurement problems. Nevertheless, measurement problems cannot be ignored in practice, and they undoubtedly exclude many household decisions from the domain of attractive applications of the household production function approach.[23]

References

Amemiya, Takeshi. "The Nonlinear Two-Stage Estimator." *J. Econometrics* 2 (July 1974): 105–10.

———. "The Nonlinear Limited Information Maximum Likelihood Estimator and the Modified Nonlinear Two-Stage Least Squares Estimator." *J. Econometrics* 3 (November 1975): 375–86.

---

restrictions cannot hinder identification and usually help considerably. See Koopmans, Rubin, and Leipnik (1950, p. 94) and Wegge (1965). Hence our results are very conservative. Also observe that only system (4) depends upon tastes. Hence, if an identification problem should arise in a particular application (although I have been unable to construct such an example), further restrictions can be introduced into the system by postulating systematic taste or technological change. The literature on the new home economics is rich in household characteristics that can be appropriate choices as shift variables in tastes and technology.

[22] A rigorous proof in the case of nonlinearity in the variables is available in Fisher (1966, pp. 148–51).

[23] Furthermore, I agree with Pollak and Wachter that some applications of that approach have (unnecessarily) imputed causation to shadow prices. But I do not believe that an approach should be rejected by identifying it definitionally with its abuses. The solution to the abuse of an approach is the proper use of the approach

———. "The Maximum Likelihood and the Nonlinear Three-Stage Least Squares Estimator in the General Nonlinear Simultaneous Equation Model." *Econometrica* (1977), in press.

Barnett, William A. "Maximum Likelihood and Iterated Aitken Estimation of Nonlinear Systems of Equations." *J. American Statis. Assoc.* 71 (June 1976): 354–60.

———. "Recursive Subaggregation and a Generalized Hypocycloidal Demand Model." *Econometrica* (1977), in press.

———. *Consumer Demand and Labor Supply.* Studies in Mathematical and Managerial Economics, edited by H. Theil. Amsterdam: North-Holland, 1978, in press.

Fisher, F. M. *The Identification Problem in Econometrics.* New York: McGraw-Hill, 1966.

Gallant, A. R. "Three-Stage Least-Squares Estimation for a System of Simultaneous, Nonlinear, Implicit Equations." *J. Econometrics* 5 (January 1977): 71–88.

Hall, Robert E. "The Specification of Technology with Several Kinds of Output." *J.P.E.* 81, no. 4 (July/August 1973): 878–92.

Koopmans, T. C.; Rubin, H.; and Leipnik, R. B. "Measuring the Equation Systems of Dynamic Economics." In *Statistical Inference in Dynamic Economic Models*, edited by T. C. Koopmans. Cowles Commission Monograph 10. New York: Wiley, 1950.

Pollak, Robert A., and Wachter, Michael L. "The Relevance of the Household Production Function and Its Implications for the Allocation of Time." *J.P.E.* 83, no. 2 (April 1975): 255–77.

Wegge, Leon L. "Identifiability Criteria for a System of Equations as a Whole." *Australian J. Statis.* 7 (November 1965): 67–77.

Reply: "Pollak and Wachter on the Household Production Approach"

Robert A. Pollak and Michael L. Wachter

University of Pennsylvania

In his comment on our paper, "The Relevance of the Household Production Function and Its Implications for the Allocation of Time" (Pollak and Wachter 1975), William Barnett (1977) raises three general issues. They involve the following: (1) the role of commodity shadow prices, (2) the identification of structural parameters when the household's technology exhibits joint production, and (3) the proper estimation of the structure. We deal with them sequentially below.

We said and Barnett agrees that there are serious conceptual problems in distinguishing between utility and production processes. In addition, there are important measurement problems in the variables that appear in the empirical household production literature. Barnett does not raise any issues connected with our discussion of these points, nor does he offer any resolutions of the problems. Hence, for the purposes of this reply, we shall avoid these issues and assume that the outputs of the relevant household production processes can be directly observed and measured and that the data on these outputs are available.[1]

On the issue of commodity shadow prices, Barnett agrees with us that they have no explanatory power as independent variables in the commodity demand functions. Barnett, however, attempts to rescue the formalism of the commodity shadow-price model by observing that the optimal solution to the household's production-consumption problem

---

The authors are professors of economics at the University of Pennsylvania. This note and the original paper were supported by grants from the National Science Foundation and the National Institute of Health and Child Development. Paul J. Taubman and Terence J. Wales provided valuable comments.

[1] In dealing with Barnett's comment on commodity shadow prices, it is useful to distinguish between the "household production approach" and the "commodity shadow-price approach." The former postulates that the household may be viewed as combining "market goods" and time to produce "basic commodities" which are the arguments of the household's utility function. The latter asserts that commodity shadow prices are legitimate and useful explanatory variables in the commodity demand functions. Although our paper did not utilize this terminology, we endorsed what we now label the household production function approach, subject to reservations about the measurement of the household's output of commodities, but reject the commodity shadow-price approach.

[*Journal of Political Economy*, 1977, vol 85, no 5]
© 1977 by The University of Chicago All rights reserved

implies a set of commodity shadow prices, π^*. He asserts that the "implicit demand function" $G(Z, \pi^*, \mu)$ has neoclassical properties and interprets this construct as an attack on our paper. We disagree that this rescues the commodity shadow-price approach. With well-behaved preferences and a convex feasible set—whether generated by monopsony, tie-in sales, or joint production in the household's technology—there does exist a hyperplane which separates the feasible set from the highest attainable indifference curve. In this sense, Barnett had definitionally solved his problem. The real problem, however, is not with the function G but with the endogenous commodity shadow prices which are its arguments. We believe that it is best to dispense with the semblance of traditional demand functions, that is, demand functions defined on commodity shadow prices, in the household production approach.[2]

On the second issue, we agree with Barnett that joint production does not preclude identification. However, we did not assert nor do we believe the contrary. We specifically state that it is possible to estimate the household's technology even when there is joint production.[3] Thus we find Barnett's entire discussion of identification irrelevant to the issues we raised in our paper.

The final issue involves the problem of estimation. Barnett argues that we incorrectly recommend estimation of a reduced form, whereas he proposes a simultaneous, structural-equation system. However, he has misunderstood our approach.

Barnett equates the commodity shadow-price approach with estimation of the structural parameters making use of the equation system

$$G[Z, \pi(P, Z), \mu] = 0, \tag{3}$$

$$X = F(Z, P), \tag{4}$$

which he calls the household's "structural form." He distinguishes this from an alternative approach to estimation based on the equation system

$$Z = f(P, \mu), \tag{5'}$$

$$X = h(P, \mu), \tag{5''}$$

---

[2] Barnett's technique could be applied to monopsony and tie-in sales as well as joint production in the household production framework. It "solves" any one of these problems only as well as it solves the others.

[3] To quote our paper, "Although the estimation of the household's technology is complicated by joint production, it is still possible" (p. 265), and "even with joint production or nonconstant returns, it is possible to estimate the household's technology and construct commodity prices" (p. 272). We do assert that joint production complicates estimation, but this is a distinct issue. E.g., in the absence of joint production and with a sufficiently simple error structure, one might attempt to estimate the production functions for each commodity separately; with joint production, the estimation of the technology cannot be decomposed in this way. In addition, joint production may involve a specification of the technology and error structure with more parameters.

which he calls the household's "reduced form." Furthermore, he asserts that we recommend estimation of (5). We object on several counts.

To see what is involved in the estimation issue, let us assume that the household's preferences belong to a class of preference orderings characterized by a vector of M independent parameters $\alpha = (\alpha_1, \ldots, \alpha_M)$ belonging to a set A (e.g., the Cobb-Douglas utility function) and that the household's technology belongs to a class of technologies characterized by N independent parameters $\beta = (\beta_1, \ldots, \beta_N)$ belonging to a set B (e.g., the Hybrid Diewert).

The investigator's task is to determine the values of the $M + N$ structural parameters, (α, β), from observations of the P's and μ's, which are exogenous, and the X's and Z's, which are endogenous. To recognize explicitly the role of the structural parameters in the behavioral relationships implied by the household production model, we write the household's cost function as $C(P, Z; \beta)$ and the commodity shadow-price function as $\pi(P, Z; \beta)$. Similarly, we rewrite Barnett's structural equations as $G[Z, \pi(P, Z; \beta), \mu, \alpha]$ and $F(Z, P; \beta)$ and his reduced-form equations as $f(P, \mu; \alpha, \beta)$ and $h(P, \mu; \alpha, \beta)$.

Contrary to Barnett's assertion, we did not suggest estimating the structural parameters by estimating the reduced-form system[4] (5). An estimation procedure more nearly consistent with what we suggest in the final section of our paper (pp. 272–73) involves the equation system[5]

$$\begin{aligned} Z &= f(P, \mu; \alpha, \beta), \\ X &= F(Z, P; \beta), \end{aligned} \tag{$*$}$$

where F denotes factor demand functions. Estimates of the underlying structural parameters (α, β) can be obtained by estimating either his ([3], [4]) or our (*) or any number of other possibilities. In particular, all of the a priori information concerning the structure of the preferences and technology are included in (*) as well as in ([3], [4]) and appear as restrictions on the values of (α, β). We were not specific about the "correct" equation system because, in the empirical context of the household production literature, it depends upon the particular problem being investigated.[6]

---

[4] Barnett's confusion may have arisen because we suggest three alternative ways of estimating the household's technology: (1) directly, by using data on inputs and outputs; (2) by estimating the cost function, using data on input prices, output quantities, and total cost; or (3) by using the implied factor demand function. It is not obvious to us that Barnett's approach, utilizing the factor demand functions, is superior to (or more "structural" than) the others.

[5] If explicit closed-form solutions for Z and X cannot be obtained, it is possible to write these equations in implicit form. In traditional demand theory, it is seldom possible to obtain explicit closed-form solutions for the demand functions $X = h(P, \mu)$ by maximizing a utility function $U(X)$ subject to the budget constraint.

Barnett correctly points out that simultaneous estimation of all of the parameters of the household's tastes and technology is desirable on econometric grounds, and he recommends a full information maximum likelihood approach. Practitioners of the household production approach, however, should derive little comfort from Barnett's econometric proposal. Restricting the household production model to situations in which estimation of the structural parameters of complete systems is possible severely reduces its usefulness as a framework for empirical work. A formulation of the household production approach which does not rely on commodity demand functions defined on commodity shadow prices can be a useful basis for empirical research. Some compromise is therefore needed between the current practice of single-equation estimation of commodities as a function of commodity shadow prices and Barnett's complete system approach.

References

Barnett, William A. "Pollak and Wachter on the Household Production Function Approach." *J.P.E.* 85, no. 5 (October 1977): 1073–82.

Pollak, Robert A., and Wachter, Michael L. "The Relevance of the Household Production Function and Its Implications for the Allocation of Time." *J.P.E.* 83, no. 2 (April 1975): 255-77.

---

[6] In our paper, we ignored the problem of stochastic specification. Incorporating a plausible error structure into either ([3], [4]) or (*) will be complex. E.g., a stochastic specification which reflects the fact that the budget constraint $\sum p_k x_k = \mu$ holds as an identity must be fairly complicated.

Book Reviews

Money, Employment and Inflation. By ROBERT J. BARRO and HERSCHEL I. GROSSMAN. Cambridge: Cambridge University Press, 1976. Pp. 264. $13.50.

Traditionally, the beginning economics student learns that equality of supply and demand is the central requirement of economic activity and that actions required to achieve this equality are what generate observable economic outcomes. At the broadest macroeconomic level, the operational relevance of this requirement follows immediately from a twin presumption of insatiability of the relevant set of desires and limitations on at least some of the resources available to meet them. Once equipped with this essential notion, the student of macroeconomics traditionally proceeds to consider the nature of the supplies of and demands for several different ingredients of an idealized economy—one or more kinds each of consumer goods, labor services, and assets—in such a way as to achieve two objectives: First, from the standpoint of what goes into them, *each* of these various supply and demand relationships should be plausible in terms of whatever economists know (or think they know) about human behavior and physical technology. Second, from the standpoint of what comes out of them, the joint implications of *all* of these supply and demand relationships should be consistent with the patterns of macroeconomic outcomes actually observed. Both objectives are relevant to Robert Barro and Herschel Grossman's thought-stimulating book.

The main contribution of Barro and Grossman's work lies in the attempt to devise a macroeconomic system which, taken on its own terms, is capable of generating outcomes corresponding to observed depressions (or recessions) of economic activity. Devising plausible macroeconomic supply and demand functions capable of generating, via general market-clearing equilibrium, underemployment outcomes similar to those often observed in large industrial economies has proved a difficult undertaking. Imaginative economists have postulated a variety of functional shapes and shifts intended to meet this objective, but still the net result has been unsatisfactory. Against this background, Patinkin (1949, 1965) working from one direction and Clower (1965) working from another suggested that economists might be making the task unnecessarily difficult by imposing the requirement of market-clearing equilibrium. Exploiting the distinction between "notional" and "effective" supplies and demands, as developed by Clower and subsequently Leijonhufvud (1968), Barro and Grossman (1971) used an apparatus similar to Mundell's (1964) to formulate a model of general market disequilibrium by symmetrically pairing Patinkin's treatment of sales-constrained firms with Clower's treatment of employment-constrained households. Indeed, the symmetry of the resulting model's construction would be even more pronounced but for the (artificial) assumption that households hold money balances while firms do not.

Chapter 2 of Barro and Grossman's book, which the authors rightly identify as the book's "analytical core," elaborates in detail the workings of this general disequilibrium model. Based on an underlying Patinkinesque general market-clearing equilibrium model developed in chapter 1, the general disequilibrium model shows how, given inflexible prices and/or wages, either excess supply or excess demand can simultaneously prevail in both the goods market and the labor market. In the simultaneous excess supply case, the model has substantial appeal as a device for understanding protracted downturns in economic activity. Here the primary result is that *all* exchange ratios must be at their "correct" values for full-employment equilibrium to prevail. Hence "too high" a real wage rate is a sufficient, but not a necessary, condition for underemployment, and the mechanics of the model bear out fully the analytical basis for Keynes's (1936) objection to wage reductions as possibly counterproductive during depression.

In addition, the general disequilibrium model exploits the principle of "voluntary exchange" (when supply and demand differ and price cannot adjust to equate them, the amount traded equals the smaller of the two) to establish an implication of this result which, in contrast to the symmetry of the model's construction, is crucially asymmetrical. In particular, *any* deviation from the price-wage vector associated with general market-clearing equilibrium will lead to *under*employment of labor and *under*production of output. Stated conversely, the implied proposition is that, given a fixed price-wage vector which would clear all markets in the absence of exogenous disturbances, *either increases or decreases* of government purchases will cause employment and output to *fall*, as will *either increases or decreases* of the money stock. This implication is somewhat startling, and it in turn bears yet further intriguing implications not only for judging the prospective use of monetary and fiscal policies but also for empirically testing the effectiveness of these policies in the past.

In view of (1) the theoretical appeal of marrying the symmetrical propositions of Patinkin and Clower, (2) the potential empirical applicability of the model's simultaneous excess supply case, and (3) the unusual nature of the resulting asymmetry implication for policy, it is not surprising that Barro and Grossman's general disequilibrium model has attracted substantial attention during the past several years.

The remaining five chapters of the book take up additional topics, primarily developing them within the context of the general market-clearing model of chapter 1 but also, along the way, showing that it is possible to append them to the general disequilibrium model too. In each case the most interesting and thought-provoking parts of the analysis are those relating to the model's dynamics. Chapter 3, for example, incorporates a treatment of financial assets and investment. Here the principal value added over Patinkin's model is to show that, by allowing for the effects of asset stocks in shifting schedules over time in the manner of Blinder and Solow (1973) and Tobin (1976), it is possible to represent the general disequilibrium model in a form analogous to an IS-LM model. Chapter 4, which relies entirely on the general market-clearing model, adds a partial-adjustment representation of goods price movements together with either exogenous or adaptive expectations of the resulting price change. Here the short-run effect of money on real and nominal asset yields is as Mundell (1963) argued, while the long-run non-superneutrality of money is as in Tobin (1965) or Foley and Sidrauski (1970). Chapter 5 adds a partial-adjustment representation of nominal wage movements, augmented by expectations after Phelps (1967) and Friedman (1968). Here the principal result is that, under specified conditions, the general disequilibrium model reproduces the familiar combination of a nonvertical short-run and vertical long-run Phillips curve, as well as elliptical movements in Phillips curve $([(1/W)(dW/dt)], U)$ space in either Lipsey's (1960) clockwise direction or Phelps's counterclockwise direction. Chapter 6 adds partial-adjustment representations of the short-run demands for money and for goods (toward the respective long-run demands) and a partial-adjustment representation of asset yield movements (thereby introducing the possibility that the asset market does not clear).

What is the net, however, of all of this further elaboration of the basic model? Given a model of general disequilibrium in the sense of the nonnecessity of market clearing, the most important positive and normative questions seem to be these: First, from a positive viewpoint, since the analytical replacement for the standard requirement of supply-demand equality is some presumed sluggishness of prices and/or wages and/or yields, what is the character of these all-important dynamic adjustments? As an implication of these adjustments, is there any tendency for the system to return to equilibrium? More specifically, as Leijonhufvud has put the question in the Keynesian context of unutilized labor resources, is the disequilibrium generated by the model temporary, or is it a true "underemployment equilibrium"? Second, from a normative viewpoint, given whatever knowledge is available in answer to these and other positive questions, what is the correct response (if any) of monetary and/or fiscal policy if such a general disequilibrium situation exists in sufficient intensity to render its protracted continuation socially unacceptable?

While the general disequilibrium model which Barro and Grossman present in chapter 2 provides a useful conceptual framework for asking these questions, the further elaboration of the model in the remainder of the book does not provide the answers. The fundamental stumbling point is that—in company with economists in general—Barro and Grossman have not solved the problem of generalized price dynamics. Throughout the book, therefore, they analyze side by side both their Patinkinesque general equilibrium model, in which all markets always clear and prices and wages and yields are always at their equilibrium values, and their general disequilibrium model, in which some price or wage or yield is fixed ad hoc and consequently some markets do not clear. They provide no way to put these two models together. In the general disequilibrium model itself, for example, economic agents who need to form expectations about the future simply assume that the currently prevailing disequilibrium situation will persist for an exogenous $\hat{N}$ of the N years in their planning horizon and that equilibrium will prevail thereafter. The prevalent dichotomization of the book's contents prevents assembling a consistent model to use in assessing the overall evolution of the macroeconomic system either on its own or in the presence of external policy stimuli.

Why does this problem of dichotomization arise? This reviewer's diagnosis is that, of the two objectives of macroeconomic model building cited above—individual theoretical plausibility of the model's component relations and their collective consistency with frequently observed outcomes—Barro and Grossman have made more noticeable progress in the direction of achieving the second. Indeed, it is the ability to describe economic depressions which, on the whole, constitutes the primary interest of the general disequilibrium model. In the book's introduction, however, Barro and Grossman also explicitly acknowledge the first objective—that is, to present ". . . a comprehensive choice-theoretic analysis of the determination of the level of unemployment and the rate of inflation . . . to explain the critical aspects of individual behavior in terms of the motivation of the individuals involved and the constraints which they face." It is in this direction, the attempt ". . . to rework the theory of macroeconomic relations through a reexamination of their microeconomic foundations," that their book is somewhat less successful. More important, from the standpoint of implications for further research, it is probably through deriving key macroeconomic relations more closely from underlying microeconomics that future work will be able to break the fundamental equilibrium/disequilibrium dichotomy of Barro and Grossman's two models.

Barro and Grossman do introduce some of the macroeconomic relations in their model, particularly those describing the behavior of households, by considering the underlying utility maximization process. The net importance of much of this analysis for the model, however, is not readily apparent. The labor supply, goods demand, and money demand functions in the general equilibrium model of chapter 1, for example, are nearly identical with those in Patinkin's model. (Since nonwage wealth Ω usually varies only with price level P in much of the analysis, most readers familiar with Patinkin would have found the exposition easier to follow if the relevant axes of the diagrams were consistently labeled with P, instead of Ω in some places and P in others.) Furthermore, in the disequilibrium context of chapter 2, the explicit treatment of consumption along the lines of the life-cycle model requires, once again, that households rely on (exogenous) expectations that the disequilibrium will persist for $\bar{N}$ more years and then give way to equilibrium; given these expectations, the resulting expressions for constrained household behavior follow in a straightforwardly analogous way. Once nonmoney assets are introduced in chapter 3, the money demand function also includes a transactions cost term γ as in Baumol (1952) and Tobin (1956), and an additional equation determines the optimal transactions frequency v; but γ and v are largely excess baggage after being introduced, playing little or no further role in the model. By contrast, the treatment of investment behavior by firms is substantially less thorough.

Most important, however, little or no "micro foundation" underlies the limited adjustability of prices and wages—which is, after all, crucial to the existence of the general disequilibrium in the first place. Furthermore, the various partial-adjustment models which Barro and Grossman use to describe the dynamics of prices and wages (and asset yields) are largely inconsistent with the construction of the general disequilibrium model itself. The partial-adjustment equations introduced at the end of chapter 2 relate the time rates of change of prices and wages to the differences between effective demand and effective supply in the goods market and the labor market, respectively, and in chapter 6 an analogous equation relates the time rate of change of asset yields to the difference between effective demand and effective supply in the asset market. If effective supply and demand in any market differ (for more than an unobservable instant), however, then what is the meaning of the "constraints" subject to which the general disequilibrium model supposes that market participants behave? Conversely, an alternative set of partial-adjustment equations for prices and wages, introduced in chapter 4, assumes that the time rates of change of these variables depend on the relevant notional demands and supplies; but a fundamental motivation for the general disequilibrium model is that it is effective, rather than notional, demands and supplies which determine actual market behavior.

The dynamic relations specifying the determination of prices and wages are the elements of the model for which careful investigation of the underlying microeconomics, specifying exactly who is setting exchange ratios, and with what objectives and under what constraints, would probably be most beneficial. Just as Barro and Grossman repeatedly emphasize that trades do take place in the absence of a Walrasian auctioneer (and without recontracting), it is essential to recognize that exchange ratios are somehow set in the same process. Price- and wage-setting behavior in a non-Walrasian context is now the object of widespread research, and the largely unstated message of Barro and Grossman's book is that progress on this front is necessary before economists can answer the key positive and normative questions posed above about the working of a general disequilibrium system.

This reviewer's conjecture is that it will prove essential to incorporate into such price- and wage-setting models the informational implications of some apparently fairly mundane mechanical aspects of the workings of non-Walrasian markets. For example, do employment-constrained workers learn that jobs are unavailable by applying at personnel offices, in which case firms acquire information about labor supply conditions, or by reading newspaper want ads, in which case firms gain no incremental information? What are the implications for price-setting behavior of the way in which firms learn that they are sales constrained? Here, as in many other aspects of the analysis, extending the model to permit inventory holdings is probably an essential next step. In addition, once firms may hold inventories, the role of the credit markets becomes central to the meaningfulness of either sales-constrained or employment-constrained behavior since, in the absence of credit market constraints, firms may produce beyond sales and households may consume beyond income. Can credit market institutions then also play a role in transferring information among firms and households?

Apart from the specifics of any one model of price and wage dynamics, it seems clear that the way to push beyond the insightful model of Barro and Grossman's chapter 2—that is, the way to integrate this general disequilibrium model with the corresponding general equilibrium model—is to derive a satisfactory model of limited-flexibility price and wage adjustment. Seen from one perspective, this task amounts to focusing on micro foundations; seen from another, it amounts to capturing macro dynamics—but, at least in this case, the two are the same.

BENJAMIN M. FRIEDMAN

Harvard University

References

Barro, Robert J., and Grossman, Herschel I. "A General Disequilibrium Model of Income and Employment." *A.E.R.* 61 (March 1971): 82–93.

Baumol, William J. "The Transactions Demand for Cash: An Inventory Theoretic Approach." *Q.J.E.* 66 (November 1952): 545–56.

Blinder, Alan S., and Solow, Robert M. "Does Fiscal Policy Matter?" *J. Public Econ.* 2 (November 1973): 318–37.

Clower, Robert W. "The Keynesian Counterrevolution: A Theoretical Appraisal." In *The Theory of Interest Rates*, edited by F. Hahn and F. Brechling. London: Macmillan, 1965.

Foley, Duncan K., and Sidrauski, Miguel. "Portfolio Choice, Investment and Growth." *A.E.R.* 60 (March 1970): 44–63.

Friedman, Milton. "The Role of Monetary Policy." *A.E.R.* 58 (March 1968): 1–17.

Keynes, John Maynard. *The General Theory of Employment Interest and Money.* New York: Harcourt, Brace & World, 1936.

Leijonhufvud, Axel. *On Keynesian Economics and the Economics of Keynes.* New York: Oxford Univ. Press, 1968.

Lipsey, Richard G. "The Relation between Unemployment and the Rate of Change of Money Wage Rates in the United Kingdom, 1862–1957: A Further Analysis." *Economica* 27 (February 1960): 1–31.

Mundell, Robert. "Inflation and Real Interest." *J.P.E.* 71, no. 3 (June 1963): 280–83.

———. "An Exposition of Some Subtleties in the Keynesian System." *Weltwirtschaftliches Archiv* 93, no. 2 (1964): 301–12.

Patinkin, Don. "Involuntary Unemployment and the Keynesian Supply Function." *Econ. J.* 59 (September 1949): 360–83.

———. *Money, Interest and Prices.* 2d ed. New York: Harper & Row, 1965.
Phelps, Edmund S. "Phillips Curves, Expectations of Inflation, and Optimal Unemployment over Time." *Economica* 34 (August 1967): 254–81.
Tobin, James. "The Interest Elasticity of Transactions Demand for Cash." *Rev. Econ. and Stats.* 38 (August 1956): 241–47.
———. "Money and Economic Growth." *Econometrica* 33 (October 1965): 671–684.
———. "Long Run Effects of Fiscal and Monetary Actions on Aggregate Demand." In *Monetarism*, edited by J. Stein. Amsterdam: North-Holland, 1976.

Public Assistance to Industry: Protection and Subsidies in Britain and Germany. Edited by W. M. CORDEN and GERHARD FELS.
London: Macmillan Co., 1976. Pp. xxv+233. £10.00.

This book presents six research studies containing basic information about the ways in which the governments of the Federal Republic of Germany and the United Kingdom have assisted or protected their industries. An introductory essay by Corden and Fels summarizes the studies, and the book concludes with a discussion by Corden of the logic of government intervention, recent developments in Britain, and various policies to accomplish the same goals at lower cost.

Hiemenz and Rabenau calculate effective protection for 34 German industries, concluding that raw material and/or labor-intensive industries benefit most from discrimination against foreign suppliers. They note that this discrimination limits the access of developing nations to German markets, because these nations are typically the main competitors in raw material and labor-intensive products.

Olton determines rates of effective protection for Britain before she joined the EEC. He tests several hypotheses relating to the structure of British tariffs and finds that it cannot be regarded in any sense as a logical system of indirect subsidization of British industry. Rather, the tariff structure of the 1960s is a collection of tariffs which were imposed at various times for various reasons and had ceased to be very relevant to the problems of the British economy at that time.

In the next chapter Fels echoes Olton's appraisal of the nature of British protection in the German context. He evaluates German assistance policy in both a qualitative and a statistical manner and allows for both tariff and various nontariff assistance mechanisms. He suggests replacing "survival assistance" with "restructuring assistance" and argues that to make economic integration politically feasible incentives must be used to move resources out of protected industries before the level of protection can be cut. Relatively unskilled *Gastarbeiter* cause diseconomies in the form of social and human problems, and German white-collar workers are in danger of becoming unemployed. Thus, he argues, incentives should be developed to divert investment so as to bring about a higher relative demand for trained and qualified labor, which would complement the effects of more restrictive immigration laws and market mechanisms in stimulating such a restructuring.

While Fels makes some attempt to rank alternative policies, I feel that he overlooked certain possibilities and could have been more critical of the government's welfare function which, for the most part, he took as given. For example, he writes that the objective of regional promotion programs "is to provide a more equal distribution of income and living standards in all parts of the country." He suggests that a superior policy would be to replace the subsidy on physical capital with one on human capital, which may or may not be true. But if the government's welfare function is as stated, and interregional migration is ignored, the best policy would be to subsidize all labor income in the depressed regions with income raised by taxes elsewhere. Clearly, this meets the stated goal with less deadweight loss than "the investment premia" used to do the job. However, it seems to me that a superior principle would be to tax equals equally and use regional incentives to overcome true externalities like congestion or pollution which is more severe in some regions than others, and even this is appropriate only if congestion or pollution cannot be taxed at appropriate rates, which would vary from region to region.

Similarly, he cites the slowness of relative wages to alter as a justification for other interventions without mentioning that the best policy would be to stimulate relative wage flexibility. It is also possible to devise mechanisms which deal more imaginatively with the political constraints in order to achieve greater efficiency. For example, when protection is a response to labor's demands and labor in an industry is successful in its bid for protection, it would be desirable to tax all capital in the industry and all further immigration of labor into the protected industry at the effective rate of protection accorded to that industry and distribute enough of these revenues to the labor that was already there to leave its wage unchanged. This would reduce the trade distortion, increase the revenue of the treasury by increasing both tax and tariff revenue, and encourage the substitution of old labor for capital. In the longer run, this would help the industry to decline with the natural attrition of its old labor force, thereby redressing the political balance so that eventually the political strength of consumers, capitalists, and new labor in the industry would dominate the strength of old labor. At this time it would become politically feasible to simultaneously dispense with the protection and the tax. Of course, the difficulty with any kind of adjustment incentives or special subsidies is that individuals have an incentive to move into industries or stay there if they anticipate being paid to subsequently leave. But this proposal seems to dominate protection by itself.

Denton gives a comprehensive account of the various interventions by government in British industry since 1960. His conclusion is that an implicit political decision has been made to prevent unemployment and hence to use labor mobility to reallocate production in light of changing comparative advantage. He cites its costs in terms of output, external payments, inflation, and, in the long term, employment itself. He notes that the incentives to efficiency and to refrain from making, or to resist, inflationary wage demands are removed if it is assumed that governments will always rescue firms that are in trouble. Finally, he notes that as the overall allocation of resources deteriorates the ultimate cost in terms of employment required to restore rational resource allocation must increase.

The last two studies compare protection of the German and British textile industries. Neu notes that German textiles are protected primarily by a quota which limits expansion by developing nations into the German market; Mills presents an interesting case study of adjustment assistance in her discussion of the varied techniques used to help the British textile industry adjust to foreign competition without generating severe unemployment.

In his concluding essay, Corden views government intervention as the product of a "conservative social welfare function" which stimulates governments to act to prevent any group from taking a cut in real income. These costs might take the form of reduced real wages or the psychic cost of moving. Such actions, he notes, do not foster change and growth, and there are always political pressures to extend them. If the maintenance of employment in an industry is the target, an employment subsidy dominates a subsidy to value added, which in turn dominates

———. *Money, Interest and Prices.* 2d ed. New York: Harper & Row, 1965.

Phelps, Edmund S. "Phillips Curves, Expectations of Inflation, and Optimal Unemployment over Time." *Economica* 34 (August 1967): 254–81.

Tobin, James. "The Interest Elasticity of Transactions Demand for Cash." *Rev. Econ. and Statis.* 38 (August 1956): 241–47.

———. "Money and Economic Growth." *Econometrica* 33 (October 1965): 671–684.

———. "Long Run Effects of Fiscal and Monetary Actions on Aggregate Demand." In *Monetarism*, edited by J. Stein. Amsterdam: North-Holland, 1976.

Public Assistance to Industry: Protection and Subsidies in Britain and Germany. Edited by W. M. CORDEN and GERHARD FELS.

London: Macmillan Co., 1976. Pp. xxv + 233. £10.00.

This book presents six research studies containing basic information about the ways in which the governments of the Federal Republic of Germany and the United Kingdom have assisted or protected their industries. An introductory essay by Corden and Fels summarizes the studies, and the book concludes with a discussion by Corden of the logic of government intervention, recent developments in Britain, and various policies to accomplish the same goals at lower cost.

Hiemenz and Rabenau calculate effective protection for 34 German industries, concluding that raw material and/or labor-intensive industries benefit most from discrimination against foreign suppliers. They note that this discrimination limits the access of developing nations to German markets, because these nations are typically the main competitors in raw material and labor-intensive products.

Olton determines rates of effective protection for Britain before she joined the EEC. He tests several hypotheses relating to the structure of British tariffs and finds that it cannot be regarded in any sense as a logical system of indirect subsidization of British industry. Rather, the tariff structure of the 1960s is a collection of tariffs which were imposed at various times for various reasons and had ceased to be very relevant to the problems of the British economy at that time.

In the next chapter Fels echoes Olton's appraisal of the nature of British protection in the German context. He evaluates German assistance policy in both a qualitative and a statistical manner and allows for both tariff and various nontariff assistance mechanisms. He suggests replacing "survival assistance" with "restructuring assistance" and argues that to make economic integration politically feasible incentives must be used to move resources out of protected industries before the level of protection can be cut. Relatively unskilled *Gastarbeiter* cause diseconomies in the form of social and human problems, and German white-collar workers are in danger of becoming unemployed. Thus, he argues, incentives should be developed to divert investment so as to bring about a higher relative demand for trained and qualified labor, which would complement the effects of more restrictive immigration laws and market mechanisms in stimulating such a restructuring.

While Fels makes some attempt to rank alternative policies, I feel that he overlooked certain possibilities and could have been more critical of the government's welfare function which, for the most part, he took as given. For example, he writes that the objective of regional promotion programs "is to provide a more equal distribution of income and living standards in all parts of the country." He suggests that a superior policy would be to replace the subsidy on physical capital with one on human capital, which may or may not be true. But if the government's welfare function is as stated, and interregional migration is ignored, the best policy would be to subsidize all labor income in the depressed regions with income raised by taxes elsewhere. Clearly, this meets the stated goal with less deadweight loss than "the investment premia" used to do the job. However, it seems to me that a superior principle would be to tax equals equally and use regional incentives to overcome true externalities like congestion or pollution which is more severe in some regions than others, and even this is appropriate only if congestion or pollution cannot be taxed at appropriate rates, which would vary from region to region.

Similarly, he cites the slowness of relative wages to alter as a justification for other interventions without mentioning that the best policy would be to stimulate relative wage flexibility. It is also possible to devise mechanisms which deal more imaginatively with the political constraints in order to achieve greater efficiency. For example, when protection is a response to labor's demands and labor in an industry is successful in its bid for protection, it would be desirable to tax all capital in the industry and all further immigration of labor into the protected industry at the effective rate of protection accorded to that industry and distribute enough of these revenues to the labor that was already there to leave its wage unchanged. This would reduce the trade distortion, increase the revenue of the treasury by increasing both tax and tariff revenue, and encourage the substitution of old labor for capital. In the longer run, this would help the industry to decline with the natural attrition of its old labor force, thereby redressing the political balance so that eventually the political strength of consumers, capitalists, and new labor in the industry would dominate the strength of old labor. At this time it would become politically feasible to simultaneously dispense with the protection and the tax. Of course, the difficulty with any kind of adjustment incentives or special subsidies is that individuals have an incentive to move into industries or stay there if they anticipate being paid to subsequently leave. But this proposal seems to dominate protection by itself.

Denton gives a comprehensive account of the various interventions by government in British industry since 1960. His conclusion is that an implicit political decision has been made to prevent unemployment and hence to use labor mobility to reallocate production in light of changing comparative advantage. He cites its costs in terms of output, external payments, inflation, and, in the long term, employment itself. He notes that the incentives to efficiency and to refrain from making, or to resist, inflationary wage demands are removed if it is assumed that governments will always rescue firms that are in trouble. Finally, he notes that as the overall allocation of resources deteriorates the ultimate cost in terms of employment required to restore rational resource allocation must increase.

The last two studies compare protection of the German and British textile industries. Neu notes that German textiles are protected primarily by a quota which limits expansion by developing nations into the German market; Mills presents an interesting case study of adjustment assistance in her discussion of the varied techniques used to help the British textile industry adjust to foreign competition without generating severe unemployment.

In his concluding essay, Corden views government intervention as the product of a "conservative social welfare function" which stimulates governments to act to prevent any group from taking a cut in real income. These costs might take the form of reduced real wages or the psychic cost of moving. Such actions, he notes, do not foster change and growth, and there are always political pressures to extend them. If the maintenance of employment in an industry is the target, an employment subsidy dominates a subsidy to value added, which in turn dominates a tariff or quota. Moreover, if relief of a real income fall is the goal, a subsidy on unemployment may be superior to the subsidization of uneconomic unemployment. Similarly, the first best policy for dealing with balance-of-payments deficit is exchange-rate change. Finally, he conjectures with respect to the British auto industry: "It is possible that for historical reasons Britain has a comparative disadvantage in an activity which depends particularly on skill in industrial relations and which is notably susceptible to the effects of industrial disruption."

To summarize, these analyses of historical experience indicate that industrial assistance policy in Britain and Germany has been determined with little attention paid to first best instruments or the long-run consequences of interferences with market mechanisms.

Edward Tower

Duke University and Simon Fraser University

The Overeducated American. By Richard B. Freeman.
New York: Academic Press, 1976. Pp 218. $4.95 (paper).

The title of this book is misleading. Both trained economists and popular readers (the target audience) will almost certainly interpret it as suggesting that excessive resources have been devoted to American higher education. Freeman has a different interpretation in mind, however. He means that graduates of American universities are overeducated in the sense that their expectations concerning the value of their education have been inaccurately sanguine in the recent past. He argues further that, as these expectations are modified toward more realistic levels, the rate of college going has and will diminish, with obvious unpleasant repercussions for suppliers of higher education.

The centerpiece of the work is Freeman's by now familiar "cobweb" model of the market for higher education. High wages for graduates are alleged to elicit enrollments which overshoot the long-run steady-state supply, resulting in corresponding underenrollments in the glut part of the cycle. This model is used to analyze a number of individual markets such as that for teachers, engineers, and lawyers as well as that amorphous aggregate, the market for bachelor's degree recipients. The postwar scenarios presented by each of these analyses are quite similar, suggesting perhaps more interdependency among these markets than is compatible with the statistical treatment given them here. At any rate, that common picture reflects a vigorous postwar market in which the demand for graduates grew rapidly enough to absorb the ever-increasing proportion of a demographically growing population of college-age men and women. It then depicts a bust of unusual severity beginning in the late 1960s.

Three influences contributed significantly to this decline. The first of these was a rather dramatic slowdown in the growth of demand for college graduates. Employment in college-manpower-intensive industries, which had grown at an annual rate of 4.4 percent in the 1960s, grew between 1969 and 1974 at only 2.8 percent. The supply of graduates, on the other hand, grew quite explosively. As unprecedented numbers of graduates of the 1960s were drawn into graduate schools, the "baby boom" cohort of graduates did not reach the market until near the end of the decade. When the graduate-education bubble finally burst, the effect on the market for college-trained manpower was startling. In the three years from 1969 to 1972, the number of male college graduates seeking work relative to the total work force increased threefold. Finally education, which is itself an important user of graduates of the educational process, is subject to an accelerator-like process of unusual force. Changes in the size of the educational industry must themselves be accommodated through the use of its own output. Thus, as college enrollments declined, the market for academics, like the tip of a whip, cracked resoundingly.

While this interpretation of postwar developments in the education sector is interesting enough, the analysis and evidence are far from conclusive. This reader wishes that the book had devoted (1) more ink to the host of interesting economic issues raised by these findings and (2) less to the presentation of semidigested facts. Instead of contrasting his scenario with competing explanations for the declining educational market and developing empirical tests which differentiate among these hypotheses, his discussion veers off into peripheral issues (i.e., race and sex discrimination) which are not illuminated by the cobweb model.

A thorough analysis of the cobweb model might have eliminated some of the ambiguities which plague the text. Perhaps the most critical concerns the role of suppliers of higher education in the dynamics of the model. For, though the influence of suppliers repeatedly crops up in the verbal discussions (particularly with reference to governmental reactions to Sputnik and the procyclical results of congressional support for higher education), the model itself seems to treat enrollments as determined exclusively by the predicted value of the training offered, that is, by demanders. Clearly this is an oversimplification.

Unless universities and their supporters respond slavishly to the applicant rate and ignore any other considerations in determining capacity, the model is misspecified, and estimates and forecasts will contain bias. Freeman's model wants to convince us that the boom and bust recently experienced in higher education are the result of an inefficient algorithm for estimating expected earnings on the part of potential university students. A plausible alternative hypothesis consistent with much of the anecdotal discussion is that the cobweb exists—not on the demand side—but on the supply side of the market for higher education.

A recursive adjustment model containing both demand and supply behavior has been developed and estimated for the medical-education market (Lindsay, Hall, and Leffler 1976). It suggests that variance in enrollment here is explained predominantly by supply rather than demand influences. Surprisingly Freeman, in his lengthy discussion of medical education, does not estimate even his simple model for medical-school enrollments. Had he done so, instead of rehashing some obsolete findings of others and overstating the conclusion which may be legitimately drawn from them, he might have noticed that there is little in the postwar history of this market (at least in responsiveness to market conditions) which differentiates it significantly from those other markets which he believes function smoothly. Individual medical-school enrollments are not significantly related to their applicants but are highly responsive to "optimal capacity" defined in terms of "zero rent" stocks in the medical-care sector (Lindsay et al., chap. 7).

Other opportunities to raise and deal with issues of considerable interest and importance were missed. Perhaps the most obvious of these is how, in the face of mounting evidence on the rationality of expectations in other markets, expectations of college graduates and those obtaining advanced degrees can have erred so dramatically. Freeman's results seem to attest to the conclusion that 1968 earnings of college graduates were not the best measure of expected earnings over the career of someone entering the market in that year. Yet his analysis assumes, without apology or explanation, that this is precisely the variable employed in the decisions of these applicants. It should be pointed out that, if enrollments were influenced by educational planners with correct expectations, then cobwebs of this type would not result– even if applicants behaved as depicted.

Furthermore, in a book in which the relationship between education and earnings plays such a critical role, one might have expected greater attention to be paid to the controversy concerning whether the observed correlation reflects an actual causal connection between education and productivity or either (1) that education is a consumption good with an income elasticity greater than zero or (2) that formal education is a screening mechanism. If the former is true, then Freeman's forecasts for this market are likely to be in serious trouble; and if the latter is true, we may have overeducated Americans in spades. Neither possibility is considered.

This is not to argue that there is nothing to be learned from this book. Freeman makes quite plausible the possibility that cobwebs actually exist (or did exist before the publication of this already influential book) *somewhere* in the market for higher education. The topics raised in this book are both timely and important, and it will be widely read and influential for these reasons. This is reassuring because, listening to educational planners, one occasionally is led to wonder whether they, in the practice of their craft, recognize the implicit constraint on the ability of the economy to consume educational services posed by the demand for higher education. There is far more than demand which is of interest and concern in the area of higher education, however, and it is unfortunate that more of the behavior surrounding the dynamics of this intriguing market was not examined. It is not enough to assert that Americans in the early 1970s were overeducated, we should also like to know why Freeman's book fails to present a convincing answer to this question.

COTTON M. LINDSAY

University of California, Los Angeles

Reference

Lindsay, C. M., Hall, T. D.; and Leffler, K. B. "The Market for Medical Education." Mimeographed. Found. Res. Econ. and Educ., Los Angeles, 1976.

The International Operations of National Firms: A Study of Direct Foreign Investment.
By STEPHEN HERBERT HYMER.
Cambridge, Mass.: M.I.T. Press, 1976. $12.50.

The major theme of the late Stephen Hymer's thesis is indeed simple. Yet that theme has dominated the bulk of the literature on international direct investment and the multinational enterprise since the early 1960s. According to Kindleberger's introduction, it was the oversimplicity of Hymer's thesis that prevented its publication by the M.I.T. Press in 1960. Major ideas are often simple; the work that embodies Hymer's straightforward and simple notions has become a seminal reference in international investment.

The approach to international direct investment introduced in chapter 1 marks one of the earlier breaks with the interest rate theory of international capital. The essential difference between direct and portfolio investments, demonstrated by the existence of cross-movements of international capital within industries, casts doubt on the generality of the interest rate approach. The interest rate theory does not explain the essential element that characterizes direct investment, *control.* Control is involved in direct international investment for two main reasons. First, direct investment requires control for the "prudent use of assets" (i.e., control assists in minimizing risk in international ventures). Second, control is an integral element of *international operations* and serves either or both of two functions: (*a*) control removes competition between enterprises operating in different countries, and (*b*) control can ensure the full appropriation of returns on certain skills and abilities possessed by the investing firm. The competition/advantages view has developed into what has become the industrial organization approach to international investment. The major statement of this approach can be found in Johnson (1970) and Caves (1971). Hymer's theory of international operations, encompassing direct international investment as one form of overseas involvement, is an extension of the theory of the firm. Firms will engage in horizontal or vertical integration across national boundaries in order to maximize returns. The nature of "advantages" specific to the firm, in the context of market imperfections relevant to that firm, will determine the extent and the form of international operations. This is the Hymer thesis.

Chapter 2 elaborates the basic imperfect competition approach to international investment. A number of important features of international operations are also introduced. In particular, Hymer argues that firms engaging in international operations must possess "advantages," such as lower-cost factors; know-how skills; distributional and marketing advantages; expertise in product differentiation in order to overcome the various barriers to operating in foreign markets, such as information costs, exchange-rate risks; and discrimination against foreigners. A further novel concept which has received some attention in recent literature (Buckley and Casson [1976] and Magee [1977]) is that direct investment as one form of international operations (compared with licensing or exports) takes place in order for the firm to *internalize* or supercede the market where the market operates less than perfectly.

The following two chapters extend the basic argument with some implications of the international operations approach (chap. 3) and limited empirical evidence based on U.S. international direct investment (chap. 4). It is of some interest to compare certain of the conclusions reached in chapter 4, based on very limited data from the later 1950s, with more recent evidence. Thus, Hymer concludes that "certain U.S. industries have significantly more foreign investment than other industries." The concentration of international investment by industries, even if the industries involved have marginally altered, is still apparent. "The industries in which there is much foreign investment tend to be concentrated industries." Foreign investment does seem to flow into the more concentrated industries (Parry 1977), though this may not be the direct result of direct investment *per se* (Knickerbocker 1976). And, "Most of the investing is done by large firms." Indeed size does emerge as a significant characteristic of U.S. direct international investment (Horst 1973).

Chapter 5 relates international financial capital movements to the theory of international operations. The main conclusion is that international money capital does respond to interest rate differentials because of the financing behavior of direct investors. Where interest rates are lower abroad, then the direct investor will tend to minimize capital transfer and borrow in the local market. Indeed a large part of U.S. international operations are financed by borrowings in local markets, a conclusion which is still generally true.

Finally, chapter 6 considers some effects of international operations on host country income and welfare. The analysis is by no means complete. The main aspects of international operations which impinge on welfare are the capital flows that are associated with international operations, the technology flows that accompany direct investment (and licensing), and the centralized control that is involved in direct investment. Interestingly, centralized control is considered from the point of view of monopoly pricing and integrated markets. It is not until some years later that Hymer considers the broader issue of potential conflict between the nation-state and the multinational enterprise: the so-called effects of interaction that circumscribe the areas of concern over the multinationals (Hymer 1970).

The Hymer thesis has been published some 16 years after its writing. The novelty of the original thesis and its importance in the development of the subject of international investment will be appreciated by anyone who has worked or intends to work in this area.

THOMAS G. PARRY

University of New South Wales

References

Buckley, P., and Casson, M. *The Future of Multinational Enterprise.* New York: Macmillan, 1976.

Caves, R. E. "International Corporations. The Industrial Economics of Foreign Investment." *Economica*, n.s. 38 (February 1971): 1–27.

Horst, T. O. "Firm and Industry Determinants of the Decision to Invest Abroad: An Empirical Study." *Rev. Econ. and Statis.* 54 (August 1973): 258–66.

Hymer, Stephen H. "The Efficiency (Contradictions) of Multinational Corporations." *A.E.R.* 60 (May 1970): 441–48.

Johnson, Harry G. "The Efficiency and Welfare Implications of the International Corporation." In *The International Corporation*, edited by C. P. Kindleberger. Cambridge, Mass.: M.I.T. Press, 1970.

Knickerbocker, F. T. "Market Structure and Market Power Consequences of Foreign Direct Investment by Multinational Corporations." Mimeographed. Center for Multinational Studies, Washington, D.C., 1976.

Magee, S. P. "Home Country Jobs and Foreign Direct Investment: Qualitative and Quantitative Impacts." In *The Economic Effects of Multinational Corporations*, edited by R. G. Hawkins. Greenwich, Conn.: JAI Press, 1977.

Parry, T. G. "Competition and Monopoly in Multinational Corporation Relations with Host Countries." In *The Economic Effects of Multinational Corporations*, edited by R. G. Hawkins. Greenwich, Conn.: JAI Press, 1977.

Leisure, Home Production, and Work—the Theory of the Allocation of Time Revisited

Reuben Gronau

The Hebrew University of Jerusalem and
National Bureau of Economic Research, Stanford

The paper tries to formalize the trichotomy of work in the market, work at home, and leisure. Time is used at home to produce home goods that are perfect substitutes for market goods, where home production is subject to diminishing marginal productivity. An increase in the market wage rate is expected to reduce work at home, while its effect on leisure and work in the market is indeterminate. An increase in income increases leisure, reduces work in the market, and leaves work at home unchanged. These conclusions are supported by empirical tests based on the Michigan Income Dynamics data, as well as by previous time budget studies. Further implications for labor supply, fertility, gain from marriage, demand for child care, and the measurement of home output are investigated.

I. Time Budget Evidence—Data in Search of a Theory

The household production function is by now an established part of economic theory. As formulated by Becker, Lancaster, Muth, and others, the new consumption theory emphasizes the fact that market goods and services are not themselves the agents which carry utility but are rather inputs in a process that generates commodities (or characteristics) which, in turn, yield utility. A second feature, introduced into the analysis by

This paper was written while I was on sabbatical at the National Bureau of Economic Research. It has not undergone the full critical review accorded the NBER studies. Research on this paper was supported by a grant to the National Bureau of Economic Research from the Rockefeller Foundation. The paper was inspired by discussions with Yoram Weiss and Robert J. Willis. I am grateful to Orley Ashenfelter, Gary Becker, Victor Fuchs, Zvi Griliches, Robert Michael, Jacob Mincer, Donald Parsons, and Shmuel Sharir for their comments on earlier drafts on this paper and to Kris Chinn and Kyle Johnson for computational assistance.

[*Journal of Political Economy*, 1977, vol. 85, no. 6]
© 1977 by The University of Chicago. All rights reserved.

and leisure. A change in the wife's wage does not affect her husband's work in the market but is positively correlated with his work at home and, as a result, negatively correlated with his leisure. An increase in the husband's wage rate increases his own supply of labor (mainly at the expense of his work at home), but reduces his wife's. This change does not affect the wife's work at home, and consequently it increases her leisure. An increase in unearned income reduces the supply of labor of both husband and wife, it reduces work at home (at least in the case of women), and it thus increases leisure.

Finally, both studies concur that children cause their mother to transfer time from the market to home tasks. However, the amount of time transferred falls short of the additional time required to care for children, so that leisure is reduced. Children have the same downward effect where the father's leisure is concerned, but in this case the father increases both work at home and work in the market.

The total time available for work at home and leisure depends to a large extent on the person's employment status. Comparing the allocation of time of employed and nonemployed Israeli married women (table 2), one can observe that, when education is controlled for, the employed have less leisure than the nonemployed. The employed Israeli married woman worked on the average 4.3 hours in the market. She saved 2.8 hours by cutting her work at home, but 1.5 hours had to come at the expense of leisure and time spent on physiological needs.

Married men work more in the market than the unmarried, and married women spend more time than the unmarried in work at home (and somewhat less in the market). Consequently, it is observed (Gronau 1976*a*) that married people have less leisure than the unmarried, and the difference is greater for men than for women. These differences are explained by two factors—marriage and the existence of children. To isolate the effect of marriage, I ran separate regressions for all men and for all women who had no young children (i.e., children in age group 0–5, or alternatively in the age group 0–12). The dependent variable is the time spent on the activity, and the explanatory variables include the person's age, schooling, continent of birth, length of residence, and number of older children; marital status is represented by a dummy variable. (For lack of space, I do not present the detailed regressions here.)

Controlling for the number of children (and the other socioeconomic variables), I found that marriage reduces the Israeli wife's supply of work to the market and increases her work at home. The decline in work in the market (about 1.5 hours a day) is somewhat smaller than the increas in work at home (about 2 hours), but the difference is too small to b significant (time spent on physiological needs and, to a lesser extent, tim spent on leisure decline, but the decline is not statistically significant)

TABLE 2

TIME-BUDGET SURVEY—AVERAGE TIME USES OF ISRAELI MARRIED WOMEN, BY EDUCATION AND EMPLOYMENT STATUS (1968)

	YEARS OF SCHOOLING											
	0–8			9–12			13+			Total		
	Employed	Not Employed	Total	Employed	Not Employed	Total	Employed	Not Employed	Total	Employed	Not Employed	Total
Age (years)	40.2	42.8	42.1	40.2	40.2	40.2	38.5	38.2	38.4	39.8	41.4	40.8
Schooling (years)	4.80	4.92	4.89	10.63	10.52	10.56	14.13	13.57	13.94	9.58	7.83	8.46
Continent of birth[a]	61	58	59	19	28	25	4	8	6	29	42	38
Number of children:												
Aged 0–5	0.46	0.74	0.66	0.41	0.68	0.58	0.44	0.70	0.53	0.43	0.71	0.61
Aged 6–12	0.98	0.93	0.94	0.24	0.51	0.42	0.34	0.49	0.39	0.51	0.73	0.65
Aged 13–17	0.85	0.90	0.89	0.38	0.35	0.36	0.34	0.35	0.34	0.53	0.64	0.60
Monthly income (IL)	670	546	578	990	843	896	1,380	915	1,221	983	694	797
Labor force participation (%)	.	. .	26	...	...	36	.	...	66	...	..	36
Time use (hours):[b]												
Total work	9.18	7.79	8.15	8.82	6.84	7.55	7.87	6.28	7.32	8.70	7.29	7.79
Work at home	4.85	7.67	6.94	4.30	6.71	5.84	3.97	6.26	4.75	4.40	7.18	6.19
Work in the market	4.34	0.12	1.21	4.52	0.13	1.71	3.90	0.02	2.57	4.30	0.11	1.60
Physiological needs	9.95	10.30	10.21	9.49	10.36	10.04	9.78	10.62	10.06	9.71	10.35	10.12
Leisure	4.63	5.27	5.10	5.16	6.14	5.78	5.86	6.50	6.08	5.16	5.71	5.52
N observations	93	265	358	117	208	325	71	37	108	281	510	791

[a] Percent of cell born in Asia-Africa

[b] Because of missing data the time uses do not always add up to 24 hours.

As for men, they hardly increase their work at home but significantly increase their supply of labor to the market (by about 2 hours). This results in a significant drop in married men's leisure.

These findings give rise to several questions: Why do education, the wage rate, and income differ in their effect on work at home and leisure? What explains the effect of children? What explains the differences in the allocation of time between labor force participants and nonparticipants? What explains the differences between men's and women's time-use patterns? How can one explain the effect of marriage, and what is the source of the asymmetry in the effect of marriage on the husband's and wife's time-use patterns? In answering these questions, we shall see that the distinction between consumption time and production time (i.e., leisure and work at home) has implications reaching far beyond the analysis of home time use, embracing such topics as labor supply, fertility, marital stability, consumption (and in particular the demand for substitutes for the person's home services), and the reevaluation of the contribution of housewives to total economic welfare.

The paper opens with a description of a theoretical model that seems to provide us with a unifying explanation of the observed time-use patterns. Some of the crucial assumptions of this model are tested in Section III. The implications of the model for the analysis of fertility, marital stability, the demand for housemaids and child care, and the evaluation of the output of the home sector are investigated in Section IV. A summarizing section discusses some qualifications and suggests some future research.

II. The Model

An intuitive distinction between work at home (i.e., home production time) and leisure (i.e., home consumption time) is that work at home (like work in the market) is something one would rather have somebody else do for one (if the cost were low enough), while it would be almost impossible to enjoy leisure through a surrogate. Thus, one regards work at home as a time use that generates services which have a close substitute in the market, while leisure has only poor market substitutes. In the extreme, though by no means unusual, case, work at home and work in the market are perfect substitutes as far as the direct utility they generate is concerned, and a person is indifferent to the composition of the goods and services he consumes, that is, to whether they are produced at home or purchased in the market.

Formally, let there be a single-person household. The person maximizes the amount of commodity Z, which is a combination of goods and services

(X) and consumption time (L),

$$Z = Z(X, L). \tag{1}$$

The goods can either be purchased in the market or produced at home, but the composition of X does not affect Z.[3] I shall measure the value of home goods and services (X_H) in terms of their market equivalents (i.e., the cost of the quality-corrected good in the market). Let X_M denote market expenditures; then total consumption is composed of the consumption of goods purchased in the market and those produced at home,

$$X = X_M + X_H. \tag{2}$$

Home goods are produced by work at home (H),

$$X_H = f(H), \tag{3}$$

subject to decreasing marginal productivity ($f' > 0, f'' < 0$).[4] The decline in the value of marginal productivity at home is due not only to fatigue or changes in input proportions but also to a change in the composition of X_H—a shift, as H increases, toward activities that have a cheaper market substitute.

The maximization of Z is bound by two constraints: the (endogenous) budget constraint

$$X_M = WN + V, \tag{4}$$

where W is the person's wage rate (assumed to be constant), N denotes market work, and V other sources of income; and the time constraint[5]

$$L + H + N = T. \tag{5}$$

The necessary conditions for an interior optimum call for the marginal product of work at home to equal the marginal rate of substitution between goods and consumption time, which in turn equals the shadow price of time, W^* (eq. [6]). If the person works in the market ($N > 0$), they

---

[3] This assumption is crucial to the model and distinguishes it from previous formulations such as $Z = Z(X_M, X_H, L)$ which had only very limited predictive power (Gronau 1973).

[4] For simplicity I ignore the market goods that enter into the production of home goods.

[5] Thus one can easily rewrite eq. (1) as

$$Z = Z(X, L) = Z'(X, L, T - L) = Z'(X, L, H + N), \tag{1a}$$

i.e., eq. (1) does not imply that work at home and work in the market do not affect welfare, but merely that H and N are perfect substitutes as far as the consumption technology (Z) is concerned.

will also equal the real wage rate, W (eq. [6a]).[6]

$$\frac{\partial Z/\partial L}{\partial Z/\partial X} = f' = W^{*}, \tag{6}$$

$$\frac{\partial Z/\partial L}{\partial Z/\partial X} = f' = W^{*} = W. \tag{6a}$$

These conditions are depicted in figure 1. The home production function is described by the concave curve $TB_0'A_0C_0$. The more time the individual spends working at home (as measured by the horizontal distance from point T), the greater the amount of home goods produced. If the individual spends all his time in work at home, he can produce an amount of OC_0 units of goods. In the absence of market opportunities, the curve $TB_0'A_0C_0$ is the opportunity frontier enclosing the set of all feasible combinations of X and L. The existence of a market where the person can sell his working time and buy market goods expands this set. Thus, given the real wage rate W (described by the slope of the line A_0E_0), the person can trade his time for goods along the price line A_0E_0 (the line tangent to the production curve $TB_0'A_0C_0$). At the optimum the person may choose a goods-intensive combination of X and L, such as B_0, where he enjoys OL_0 units of consumption time, spends L_0N time units on work in the market, and spends NT time units on work at home. Alternatively, the person may have a high preference for leisure (i.e., a leisure-intensive consumption technology), choosing as his optimum combination the point B_0'. In this case he does not work in the market, but splits his time between leisure (OL_0') and work at home ($L_0'T$).

Note that the person may adopt a goods-intensive technology, such as B_0, but it may still be home-time intensive in the sense that a large part of the goods are produced at home. Thus, leisure intensive and home-time intensive are not synonymous. Note further that if the marginal productivity of work at home at the point T falls short of the real wage rate, there is no home production and we are faced with the familiar Robbins diagram and the dichotomy of work (in the market) and leisure.

To analyze the properties of this model, let it be assumed that there is an increase in other sources of income by an amount of ΔV. An increase in other sources of income secures for the person the amount of OX_0 of

---

[6] Eqq. (6) and (6a) are derived by maximizing the Lagrangian function $G = Z\{[X_M + f(H)], L\} + \lambda(WN + V - X_M) + \mu(T - L - H - N)$ with respect to L, H, N, and X_M. The shadow price of time (measured in real terms) equals $W^* = \mu/\lambda$, where μ and λ are the marginal utilities of time and income, respectively. The wage rate may fall short of the value of marginal productivity at home ($W < f'$), either because of the person's reluctance to perform the home services outside of his own home or because of differences in the value of marginal productivity between home and the outside, due to transport costs, monitoring costs, and efficiency (the person being self-employed in his own home).

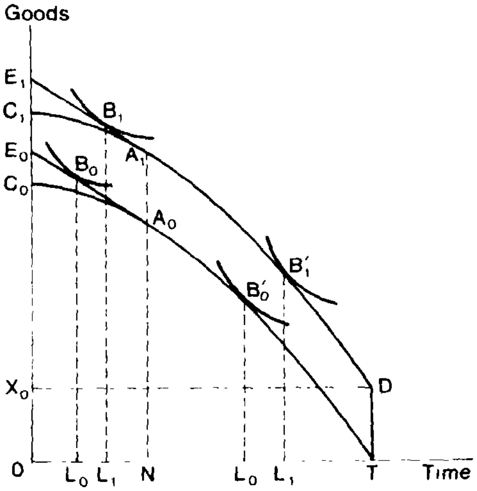


FIG. 1

market goods even if he spends all his time in consumption. The change is reflected, therefore, in a vertical shift of the production curve $TB'_0A_0C_0$ to $TDB'_1A_1C_1$. The change does not affect the marginal productivity of work at home—it does not affect the shape of the curve but only its location. Since the real wage rate is given, there is no change in the point at which the person finds it cheaper to buy the goods in the market than to produce them at home. If the person prefers a goods-intensive consumption technology which makes him work in the market (combination B_0), he does not change the amount of time he spends working at home (NT) and, given the pure income effect, he increases his amount of leisure (if leisure is not an inferior input) at the expense of work at the market (consumption time increases from OL_0 to OL_1 and work in the market is reduced from L_0N to L_1N).[7]

If, on the other hand, the person does not initially work in the market (point B'_0), the increase in income and the resulting increase in Z call for an increase in consumption time which can come only at the expense of work at home.

Let there be an increase in the real wage rate W (fig. 2). If the person works in the market (point B_0), a change in wages affects both the rate of substitution between consumption time and goods and the profitability of home production. The increase in wages lowers the price of goods in terms of time, thereby making home production less profitable and

---

[7] The decline in work in the market may result in the person's dropping out of the labor force altogether.

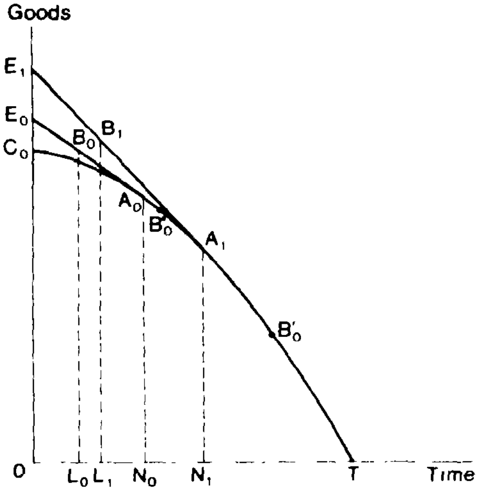


FIG. 2

inducing substitution of goods for consumption time. This change will, therefore, definitely cut work at home (from N_0T to N_1T), while its effect on leisure is indeterminate. The substitution effect tends to reduce leisure, while the expansion effect tends to raise it. As for work in the market, it depends on the extent of the reduction of work at home and on the change in consumption time. If the reduction in work at home exceeds the increase in leisure (if there is one), the supply of work to the market increases. The tendency of this supply curve to be positively sloped increases, the greater the rate of substitution between goods and consumption time, the less sensitive the marginal productivity in home production to changes in the amount of work, and the smaller the income elasticity of leisure.

If the person initially does not work, the change in wages may lure him into the market (point B_0''), or he may be completely unaffected (point B_0').

A third kind of change worth examining is a change in productivity. It is impossible to predict the implications of this change without specifying the exact nature of the changes in home productivity (i.e., changes in f) and consumption technology (i.e., changes in Z). In the absence of the necessary information, one's predictions are limited to the case where the person works in the market. In this case a change in consumption technology would affect work in the market and leisure but would leave work at home unchanged. On the other hand, an increase in the productivity of work at home is associated with an increase in real income and an

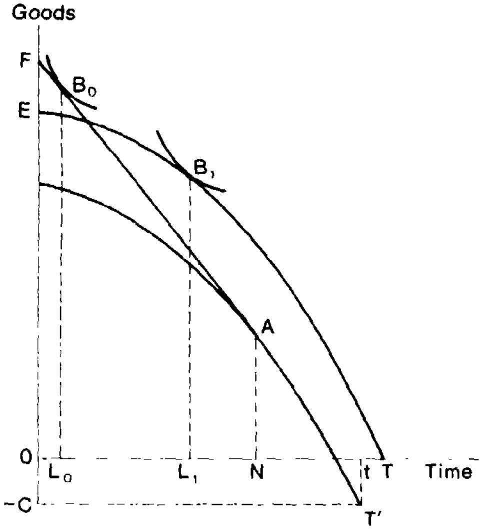


FIG. 3

increase in leisure, but its effect on work at home and work in the market is indeterminate.

Up to this point, it has been assumed that entry into the market is costless. In practice, work in the market involves costs in terms of both money and time. Let these costs be C and t, respectively, and let them be independent of the amount of work N (e.g., transportation costs and time).[8] The introduction of these costs calls for some modification of the budget and time constraints

$$X_M + \delta C = WN + V, \tag{4a}$$

$$L + H + N + \delta t = T, \tag{5a}$$

where δ is a dummy variable that describes the person's employment status

$$\delta = \begin{cases} 1 \text{ when } N > 0 \\ 0 \text{ when } N = 0. \end{cases} \tag{7}$$

The person is faced by two alternative opportunity sets (fig. 3). If he stays out of the labor force and confines himself to home production, he can choose any point on the boundary TB_1E. On the other hand, if he decides to join the labor force, he suffers a loss of t units of time and C

---

[8] Variable time and money costs (i.e., costs that vary with N) can easily be treated by an appropriate modification of the wage rate.

units of X, but his opportunity locus becomes $T'AF$. Given these opportunity sets, a person with a greater preference for goods will join the labor force (point B_0), spending OL_0 units of time on leisure, working in the market for L_0N units, working at home for Nt units, and, say, traveling to work for tT units of time. A person with a greater taste for leisure will decide to stay out of the market (point B_1), dividing his time between leisure and work at home (OL_1 and L_1T, respectively). Given the opportunity set, labor force participation is therefore associated with a decline both in leisure and in work at home.[9] The existence of entry costs does not, however, affect our previous conclusions about the effect of changes in the socioeconomic characteristics on the allocation of time.

The predictions of this simple model are by and large consistent with the reported findings. An increase in the wage rate should not affect the allocation of time of the nonemployed but should reduce the work at home of the employed. Thus, on the whole, one would expect the wage rate and work at home to be negatively correlated. The effect of a change in the wage rate on leisure depends on the relative magnitudes of the income effect and the substitution effect. The tendency for the income effect to dominate increases with the number of hours worked in the market. Thus, it is not surprising that the substitution effect is the dominant factor in the case of the wife's leisure, but the two effects cancel out (or even the income effect dominates) in the case of the husband's. An increase in nonwage income should not affect the work at home of employed persons but should reduce the work at home of the nonemployed. Consequently, one expects nonwage income and the wife's work at home to be negatively correlated. On the other hand, in the case of men, who are mostly employed, the negative effect should be much less pronounced and may be nonsignificant In either case, one expects nonwage income and leisure to be positively correlated. Finally, in the presence of market entry costs, employed persons should spend less time on work at home than the unemployed, but this difference is swamped by the difference in the market hours Consequently, one expects the employed to work longer hours (in the market and at home) and enjoy less leisure, other things being equal.

The model can easily be extended to the case of two commodities. In this case the person maximizes the welfare function $U(Z_1, Z_2)$ subject to the constraints imposed by the transformation curve between the two commodities. The latter depends on the consumption technology of each commodity (eq. [1]), the home production function (eq. [3]), and the time and budget constraints. The comparative-static properties of the extended model are very similar to those of the one presented earlier and

---

[9] An alternative interpretation of the different patterns of time use of the employed and nonemployed traces them to differences in productivity at home, the less productive person having a stronger inclination to join the labor force.

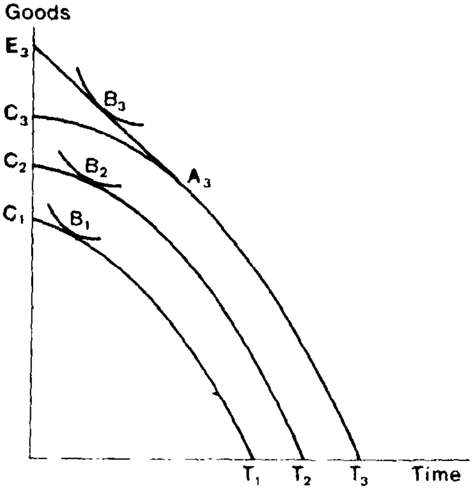


FIG. 4

will therefore not be discussed here.[10] However, the extended model sheds new light on the effect of children on the allocation of time.

An increase in the number of children (or the introduction of children) is associated with a transfer of time to child-related activities. The additional time units devoted to children are spent on work at home and leisure (a horizontal shift from $T_1B_1C_1$ to $T_2B_2C_2$ in fig. 4). Eventually the opportunities for profitable home production are exhausted, and any additional units of time spent on the commodity of children will be spent on work in the market (which will permit the purchase of child-related market goods) and leisure, leaving work at home unchanged (curve $T_3A_3B_3E_3$).

When the person participates in the labor force, the time withdrawn from other activities is initially withdrawn from work in the market and leisure, curtailing the market goods and consumption time associated with these activities. If the person is not a labor-force participant, the entire time is naturally withdrawn from work at home and leisure.

The overall effect on the allocation of time of introducing children (or of an increase in their number) depends on the leisure intensity of child-related in comparison with other activities and on the profitability of home production of child-related services. It is my contention that children are less leisure intensive (i.e., more goods intensive) than other activities. Thus an increase in the number of children at the expense of other activities should reduce the person's leisure. The effect of children on work at home and in the market depends on the profitability of home production. Women are, usually, offered lower wages than their husbands, and they may also be more productive in home production. An increase in the number of children therefore leads working mothers to work less in the market and more at home. Similarly, nonparticipants shift time from less goods-intensive to more goods-intensive activities, increasing their work at home. The scope for profitable child-related home production is much more limited for the husband. In his case one can therefore expect to observe that the shift to the more goods-intensive commodity (i.e., children) results primarily in increased work in the market (though there may also be an increase in work at home).

---

[10] For a detailed discussion see an earlier version of this paper (Gronau 1976*b*).

The price of market substitutes (maids, nursery school, kindergarten, school) declines as the child grows older. Put differently, as the child grows older, the real wage of the mother (in terms of market substitutes) increases. This increase leads to a decline in work at home and an increase in work at the market, but need not increase the demand for leisure much.[11]

Similarly, it seems that in Israel the prices of market substitutes (maids, nursery school) are cheaper than in the United States.[12] Thus, an Israeli mother should find it less profitable than an American one to divert time from work in the market to work at home when she has a child. The supply of labor of Israeli women, in particular the more educated ones, should therefore be less affected by young children than that of their American counterparts (Gronau 1976*a*).

Additional insights are gained by extending the model to the multiperson case and, in particular, the case of the married couple. Marriage introduces into people's choice set a new activity, "married life." The new activity uses home-produced goods in its production and thus involves an increase in work at home at the expense of work in the market. Furthermore, to the extent that this loosely defined activity is more goods intensive than the other activities, it should also reduce leisure. Marriage may, however, have a more fundamental effect on household members' allocation of time by allowing for specialization within the family. Much of the preceding discussion is based on the proposition that a person is reluctant (or finds it unprofitable) to sell his home goods outside the household. This assumption is, however, relaxed in the family context. Greater market involvement and discrimination result in the husband's wage rate being higher than his wife's; on the other hand, on-the-job training may make women more efficient in home production. There is therefore an incentive within the family to trade goods for time. Leisure is, by definition, an input which the person has to provide himself. Thus, there is no way in which the wife can save leisure for her husband. She can, however, save him work-at-home time. Indeed, the woman who is reluctant to sell her home goods (e.g., serve as a maid) is willing to exchange them within the family for market goods. The extent to which such an exchange takes place and the terms of the exchange depend to a large degree on her marginal costs of producing these goods.

---

[11] The goods intensity of children may decline and thus leisure may increase as the child grows older. However, as long as children are more goods intensive than other activities, they should be associated with a decline in leisure.

[12] In Israel many of the 2-year-olds and most of the 3–4-year-olds attend a nursery for at least 4 hours a day In 1968 over 40 percent of the working mothers with a child of less than 3 years old employed a maid (the fraction for working mothers with 13+ years of schooling was two-thirds).

If before marriage a woman obtained some of the goods solely through home production, the price she places on them is below that paid by her husband in the market. Specialization and trade within the family should increase the family members' welfare. The wife expands her home production while the husband increases his work in the market.[13] On the other hand, if both participate in the labor force and there is no good or service which is supplied solely though home production (i.e., $X_M > 0$ for all goods for both husband and wife), then it can be shown (Gronau 1976*b*) that before marriage both members face the same set of prices and marriage does not yield any gains of trade.

III. Some Empirical Tests

The model gives rise to a wealth of testable hypotheses. Although many of these hypotheses could have been generated also by other models, which use a weaker set of assumptions (e.g., the models discussed in my 1973 and 1976*a* papers), none of them generates this model's prediction concerning the income effect on work at home. A crucial test of our analysis focuses, therefore, on this effect: Does an increase in unearned income reduce the work at home of the nonemployed while leaving the work at home of the employed unaffected? The examination of this hypothesis is the subject of this section.

The data used are the 1972 panel of the Michigan Study of Income Dynamics. Given the peculiar characteristics of the subsample of nonemployed men (e.g., a mean age of 68), I confine the discussion to the time usage of white married women. The sample included 1,281 women, of whom 660 were employed at some time during the preceding year and 621 reported that they did not work in 1971. The dependent variables consisted of the time spent working in the market (including travel to work), the time spent in housework, and leisure.[14] The explanatory variables included the wife's age, education, and labor force experience (i.e., full-time work) since the age of 18, the husband's education and wage, the family's unearned income, the number of children below the age of 18, the number of children at school, and the number of rooms in the home. The regressions were estimated for the whole sample and separately for the employed and the nonemployed.[15]

---

[13] The wife's tendency to specialize in work at home is reinforced if work in the market involves fixed entry costs. Marriage offers the woman a job which does not involve these fixed costs at terms which may not be much inferior to her market wage rate. As a result, the wife may be tempted to drop out of the labor force and concentrate on work at home.

The findings for the whole sample closely resemble those reported earlier in Section I and are therefore not presented here. Table 3 presents the results for the two subsamples separately. The results confirm the predictions of the model. When the wife is not employed, her work at home is negatively affected (and her leisure is positively affected) by her unearned income and her husband's wage rate. Children tend to increase her home tasks, but schoolchildren less so than younger ones. Her work at home is negatively associated with her education, but positively associated with the size of her house. As predicted, her potential wage rate (as approximated by her past labor force experience) does not affect her allocation of time.

When the woman is employed, a major determinant of her allocation of time is her wage rate. This variable explains the negative effect of labor force experience on work at home and leisure and the negative effect of the wife's education on her work at home (education and leisure are positively correlated in this regression, but the regression coefficient is nonsignificant). Children have a negative effect on their mother's leisure, the time withdrawn from the market falling short of the increase in housework. As the child grows older and enters school, housework diminishes, but this change results in hardly any gains in leisure—the time saved in work at home is diverted back to the market.[16] Most important for our analysis is the income effect. The husband's wage has a significant positive effect on leisure, but no effect on work at home.

---

[14] The families reported on the number of weeks worked, the number of hours the wife worked per week, and the number of hours spent in housework in an average week. (Housework is not defined in the questionnaire, but the examples mentioned are cooking, cleaning, and other work around the house. Thus, it is not known whether the families included such activities as child care and shopping in housework.) I defined leisure as the difference between 8,760 hours per annum and the number of hours reported worked in the market and at home.

[15] Separating the sample by employment status may give rise to selectivity biases. I tried to correct for these, but had very little success because of the strong multicollinearity between the correction coefficient and the rest of the explanatory variables. It is comforting to learn that a recent attempt to correct selectivity biases in labor supply (Cogan 1976) has generated results that do not differ much from simple OLS estimates based on a sample of working wives.

[16] The effect of children on leisure, as presented in table 3, is significant at the conventional 5 percent level only if one uses a one-tailed test. However, if one removes the variable schoolchildren from the regression, the variable number of children turns out to be highly significant by any standard (i.e., *t*-values that exceed 3).

TABLE 3

The Determinants of the Allocation of Time of Married Women in the United States, by Employment Status[a]

	Not Employed		Employed					
	Work at Home		Work in the Market	Work at Home	Leisure	Work in the Market	Work at Home	Leisure
Constant	1,669.40	1,677.29	1,953.51	1,155.30	5,651.19	−9,310.70	3,213.34	14,857.36
	(6.74)	(6.69)	(8.44)	(5.32)	(21.37)	(9.14)	(3.30)	(12.61)
Wife's age	−1.165	−1.312	−15.745	3.895	11.850	−14.965	4.084	10.881
	(0.37)	(0.41)	(5.01)	(1.32)	(3.30)	(5.08)	(1.45)	(3.19)
Wife's education	−53.469	−53.811	20.740	−36.262	15.522	−241.312	11.883	229.429
	(3.28)	(3.29)	(1.36)	(2.53)	(0.89)	(8.47)	(0.44)	(6.97)
Husband's education	22.668	22.511	−32.781	14.412	18.369	−31.028	14.000	17.028
	(1.82)	(1.81)	(2.38)	(1.12)	(1.17)	(2.30)	(1.09)	(1.09)
Husband's wage ($/hour)	−16.129	−16.199	−24.017	−8.048	32.065	−151.896	15.264	136.632
	(2.21)	(2.22)	(1.80)	(0.64)	(2.11)	(8.89)	(0.93)	(6.91)
Unearned income ($10/year)	−0.441	−0.445	−0.660	0.210	0.451	−0.598	0.197	0.401
	(2.23)	(2.24)	(2.39)	(0.81)	(1.43)	(2.21)	(0.76)	(1.28)
Children aged 0–17	327.654	327.843	−198.781	278.141	−79.359	−189.559	276.025	−86.466
	(6.94)	(6.94)	(4.94)	(7.37)	(1.73)	(4.81)	(7.31)	(1.90)
Children at school	−125.196	−124.926	123.216	−104.465	−18.752	98.394	−100.098	1.704
	(2.86)	(2.85)	(3.24)	(2.93)	(0.43)	(2.64)	(2.81)	(0.04)
Rooms	83.251	83.207	6.446	27.467	−33.912	−3.502	29.148	−25.646
	(3.17)	(3.16)	(0.28)	(1.29)	(1.31)	(0.16)	(1.37)	(1.00)
Wife's experience	...	0.808	38.498	−6.244	−32.254	...	...	...
		(0.21)	(9.57)	(1.65)	(7.02)			
Wife's expected wage (log, ¢/hour)	...	...	...	...	...	2,810.95	−514.77	−2,296.18
						(11.11)	(2.13)	(7.85)
R^2	.26	.26	.19	.17	.11	.23	.17	.12
N observations	621	621	660	660	660	660	660	660

[a] Numbers in parentheses = *t*-ratios.

Similarly, work at home is not affected by changes in unearned income; their effect on leisure is positive, though weak. It is also worth noting that the work at home of employed women is not correlated with the size of their house--presumably any extra work associated with extra rooms is done by maids (or other market substitutes).

To isolate the wage effect from other effects associated with education, I introduced this variable directly into the regression. Since the survey does not include direct information on the hourly wage rate, hourly earnings are computed by dividing annual earnings by annual hours (i.e., by the product of weeks worked and weekly hours). This procedure generates serious measurement errors which bias the estimate of the wage effect. To overcome this problem, I used an indirect approach: At the first stage I estimated the wage function; at the second stage I introduced the imputed (log) wage in the time use functions. The estimated wage function was of the semilog variety, the explanatory variables being the wife's education (ED), her labor force experience (EX), and her husband's wage rate (W_h)

$$\begin{aligned}\log W = {} & -0.5955 + \underset{(9.18)}{0.0905ED} + \underset{(4.72)}{0.0302EX} - \underset{(2.99)}{0.0006(EX)^2} \\ & + \underset{(4.75)}{0.0442W_h} \\ & R^2 = .20\end{aligned}$$

where the wages are measured in dollars and the numerals in parentheses denote the corresponding t-values.

The results of the second stage (table 3) do not diverge much from our previous findings. The wage rate has a strong negative effect on both leisure and work at home. Education is positively correlated with leisure, but its effect on work at home (though positive) is not significant. (The direct effect of education on number of hours worked in the market is therefore negative.) Finally, the pure income effect is as predicted: neither unearned income nor the husband's wage rate affect the employed woman's work at home.[17] Our model also passes this test successfully.

IV. So, What's New?

The model has been shown to yield a comprehensive yet concise interpretation of the findings on the allocation of time between work in the market, work at home, and leisure. It explains the different behavior patterns of people with different incomes, wages, and education, and the effect of children on the allocation of time. It accounts for the different patterns observed for men and women, the married and unmarried, the employed and nonemployed; and it appears, on the whole, to provide the economist with more refined tools for analyzing time-budget data. But does the theory extend our understanding of household behavior beyond this goal? In this section I shall try to show that the ramifications of the theory reach far beyond the analysis of time use.

---

[17] Comparing the regression coefficients of unearned income in the work-at-home, leisure, and work-in-the-market equations, one observes that the first two are nonsignificant while the last is significant. This difference is explained by the better fit of the last equation. The difference between the husband's wage effect and the effect of unearned income is due primarily to differences in the units of measurement (dollars per hour and dollars per annum, respectively).

The Supply of Labor

The most direct application of our model is, of course, to the analysis of labor supply. In the short run the two most important economic factors affecting the supply of labor of married women are income and wage rates. According to our analysis, the income effect works primarily through its effect on leisure. On the other hand, wage increases tend to increase the supply of labor by reducing work at home, but their effect on leisure is indeterminate. Given the wage effect on leisure, the labor supply is more elastic the greater the sensitivity of work at home to changes in the wage rate (i.e., the smaller the effect of H on f'). For that reason alone, one would expect the supply of labor of married women to be more elastic than that of men. But the analysis brings up a further point: A wage increase may result not merely in a shift from work at home to work in the market but also in reduced leisure—employed women have less leisure than the nonemployed. Indeed, according to the estimates presented in table 3, the marginal effect of a wage change on the leisure of the employed is more than four times that on work at home, although the wage elasticities are almost identical (about 0.4). Admittedly, some of these changes in leisure may be due to changes in activities which are normally regarded as work at home but were not defined as housework by the respondent,[18] but it is hard to believe that this misclassification explains such a great difference.

Recent decades have witnessed a great expansion in the labor supply of married women. Still, with only about half of married women participating in the labor force, and with the number of working hours of employed women equaling the number of hours they put in on work at home, it looks as if this resource has only been partly tapped for future expansion. A natural question is to what extent will the labor-supply function of married women resemble that of their husbands once they reach similar labor force participation rates? Right now any answer to this question should be regarded as sheer speculation, since so much depends on changes in role differentiation and on reallocation of work at home within the family. It is, however, worth noting that right now women are more willing to dispense with leisure in response to wage increases than men (the change in leisure induced by a 1 percent change in wages is almost twice as great for employed wives as for their husbands) and that men are apt to increase their leisure in response to an increase in income much more than women (the ratio of the income effects is about 5:1).[19]

---

[18] See n. 14. According to the Israeli data, housework (not including child care) accounts for only two thirds of the time defined by me as work at home.

In the long run, changes in wages are associated with changes in education. But while in the short run wage increases, education given, reduce both work at home and leisure, changes in education (when the wage is not held constant) affect only work at home. The long-run prospects for increased labor supply due to increases in education are therefore much more limited. On the other hand, one can expect further expansion in labor supply if the increase in education and wages is associated with a decline in fertility.

The Demand for Children

It is customary for economists (e.g., Willis 1973) to argue that children are a home-time-intensive activity and that an increase in children therefore reduces work in the market. A corollary of this contention is that since children are time intensive, an increase in their mother's wage rate should raise their price relative to that of other commodities. Given our analysis, one has to distinguish between home-time intensity and leisure intensity (or average and marginal home-time intensity). Children may be home time intensive when they are introduced into the household, since some of the goods used in the activity can be profitably produced at home. However, as the activity increases, the profitability of home production diminishes and eventually the family relies solely on market goods (on the margin). At this point the goods-intensive nature of children becomes apparent. Thus a wage increase raises the price of children in the range where children's goods are produced at home; but when these goods are market produced, the price of children is reduced by a wage increase. The tendency to replace home goods by market goods increases with the mother's wage rate. Thus, one would expect that the price of children increases with the mother's wage for low-wage mothers but that this relationship is reversed as the mother's wage increases. Ben-Porath (1973) observes a transposed J-shaped relationship between fertility and education for Israeli women: Fertility declines with education, but there is a slight inflection at the top. These findings are consistent with the prediction of the model.[20] The price of housemaids relative to the wife's wage seems to be lower in Israel than in the United States, and the tendency to substitute housemaid services for wife's time is therefore greater in Israel. Consequently, the transposed J-shaped relationship between fertility and wife's education should be more pronounced in the Israeli data.

---

[19] The findings for the employed married men (not presented here) are based on the same sample as those for the women.

The price of market substitutes relative to the wife's wage declines as the child gets older, and the goods-intensive nature of children therefore becomes more explicit. The relative price of "older children" has a greater tendency to decline as the parent's wage increases. The overall effect of changes in the wage rate on the present value of the cost of children is therefore indeterminate and depends on the price of market substitutes, the rate of discount, etc.

Gains from Marriage

Previous studies (e.g., Becker 1973) have asserted that gains from marriage depend on the husband-wife wage ratio. Other things being equal, the higher the husband's wage rate relative to his wife's the greater the opportunity for specialization within the household and the greater the gains from trade. This conclusion must be somewhat modified if one realizes that there exists no direct way of trading leisure and that the exchange is confined to home goods. The scope for gains from exchange within the household is limited by the profitability of home production. The latter in turn depends on the wife's home productivity and the price of market substitutes. Given the wife's home productivity and the price of market substitutes, the higher the wife's wage the greater the probability that any change in her activities will not affect her home production. In this case the prices of goods confronting men and women are the same and there are no gains to be reaped from trade. The gains from trade therefore decline as the wife's wage increases irrespective of the husband's. The increase in marital instability which has accompanied the increase in women's real wage rates and their increased labor-force participation is consistent with the prediction of the model, though one does not observe any substantial narrowing of the sex wage gap (Fuchs 1974).[21]

---

[20] Ben-Porath (1973) explained this relationship in a somewhat similar fashion, arguing that if the elasticity of substitution between time and goods in the production of children exceeds unity, children may be a time-intensive commodity for low-wage mothers but a goods-intensive commodity for high-wage mothers.

[21] Fuchs (1974) reports that the sex differential in hourly earnings of white nonfarm employed hardly changed in the last decade (from 0.61 in 1959 to 0.64 in 1969). Moreover, the differential for the young (below 35) married (the group most prone to divorce) has even slightly increased (from 0 73 to 0.70).

Taxes, Child-Care Programs, and the Demand for Domestic Help

It is often claimed that the wife's entry into the labor force involves costs such as child-care and housemaid services which by far exceed the husband's cost of entry. According to this argument, child-care services should be tax deductible, as are books or other costs a person must undertake in order to work. This argument has been accepted by many legislators and incorporated into the tax laws.

The economic validity of such deductions apart, it seems that this popular argument is wrong. An increase in expenditures on child-care services is associated with the wife's work in the same way an increase in the expenditures on a gardener is associated with the husband's work. Unlike expenditures on books or commuting costs (and time), which are a prerequisite for work, they are a cost which the family willingly undertakes because it finds that it is unprofitable for the wife to spend her time in child-care activities.

The analysis emphasizes, however, an additional point. In evaluating the various child-care programs which have been proposed or enacted in recent years, one has to distinguish between their effect on the marginal rate of substitution between goods and leisure and their effect on the profitability of home production. Assume a one-commodity world where the only commodity is children. A program which gives the mother a fixed child-care subsidy for every hour worked is equivalent to a wage increase and affects both the profitability of home production and the price of leisure. On the other hand, a fixed cash rebate or free child-care services which are conditional on a minimum number of working hours do not affect the profitability of home production of working women and may only affect their demand for leisure. When it comes to nonemployed women, this kind of program encourages labor force participation (in particular if entry into the market involves fixed costs) and may affect both home production and leisure. Finally, a tax deduction for child-care expenditures which declines gradually with earnings may affect home production but need not affect the price of leisure.

It has been argued (Heckman 1974) that to evaluate and compare the impact of various programs on labor supply and welfare it is sufficient to know the indifference curves between market goods and nonmarket time. Our analysis indicates that this knowledge may not be sufficient and that a thorough evaluation may require specific information on both the household production function $f(H)$ and consumption technology (Z).[22]

Finally, it seems at first puzzling that work at home is so insensitive to changes in income, given the high income elasticity of the demand for housemaids.[23] The puzzle is solved, however, if one realizes that the demand for housemaid services (like that for many other services) is an excess demand. An increase in income does not increase the profitability of producing these services at home when a person is employed and reduces the profitability when he is not employed (the shadow price of time increasing). Thus, changes in income may have a strong effect on the excess demand for these market services and no effect (or even a negative one) on home services.

---

[22] One can easily incorporate in the analysis additional proposals (e.g., a subsidy confined to institutional forms of child care) and complicate it by introducing additional activities or exchange within the household, but this would not change our basic conclusion.

The Evaluation of Home-Sector Output

A long-standing complaint against the current national accounting system is its omission of the output of the nonmarket sector and, specifically, the output of wives at home, which according to some estimates (Morgan, Sirageldin, and Baerwaldt 1966) constitutes close to 40 percent of measured GNP. Several attempts have been made to correct this shortcoming (Morgan et al. 1966; Nordhaus and Tobin 1973; Sirageldin 1969), but these attempts were accompanied by controversy over which prices should be used to evaluate the wife's output. Should one infer the value women assign to their time and use that (Gronau 1973), or should one use the market prices of the services rendered by the wives (Walker and Gauger 1973)? The present model provides an analytical tool to resolve this controversy.

If the wife works in the market and at least some of each good is purchased ($X_{Mi} > 0$), her value of marginal productivity at home equals her wage rate ($f_i' = W$). Since most of the goods produced at home are services, she assigns to them a value that equals the wage she would have to pay somebody else to do the work for her divided by the average productivity of that person ($= W_H/AP_{Hi}$). The value placed on the last unit of work at home is therefore $W = W_H f_i'/AP_{Hi}$. Using the wage W_H of the services worker who can replace the wife in home production serves as a good approximation for the value of her time only if his or her average productivity equals the wife's marginal productivity. If $f_i' > AP_{Hi}$, the wage of the services worker will understate the price the wife assigns to her marginal unit of time in work at home.

Similarly unsatisfactory seems to be the method that uses the person's wage rate for the imputation. Women who do not participate in the labor force assign to their time a price that exceeds the wage they expect to get in the market. More important, the wage rate may serve as a close approximation to marginal, but not average, productivity at home. The product of the average wage rate and the number of hours worked at home therefore understates the value of home production to the extent that diminishing marginal productivity prevails. This imputation does not account for the rent (i.e., the producer's surplus) accruing to a person who is self-employed in his own home.

---

[23] Using Israeli data (the Family Expenditure Survey 1968–69), I found that even when one controls for the wife's education and employment status, the income elasticity of housemaids exceeds unity.

V. What Next?

I believe that this paper provides ample evidence for establishing the distinction between work at home and leisure as an integral part of the theory of the allocation of time and household production. It has been shown that this distinction is a prerequisite for any further investigation of time-use patterns and is highly useful in the analysis of fertility, marriage, child-care programs, labor force participation, and the evaluation of the output of the nonmarket sector. I am confident the model will also be found fruitful for the analysis of problems in other fields, such as medical economics or transport demand, in which the household production model has been put to good use.

It is clear that the model is incomplete I expect major criticism to be launched against the assumption that work at home involves the same marginal utility as work in the market. Child care, cooking, gardening, etc., clearly create direct utilities (positive or negative). The psychic income derived from these activities relative to that derived from work in the market may vary with the person's socioeconomic characteristics and affect his behavior. Admitting the validity of this criticism, I contend that it is not more serious than in the case of the dichotomy of home time versus work in the market. Psychic income (or leisure on the job) is an important determinant of investment in human capital, occupational choice, and the supply of labor. Economists so far have not been able to derive a satisfactory method of isolating this factor. This has not prevented research on the determinants of the supply of work to the market, and it should not block research on the supply of work at home.

A second point of criticism may focus on the neglect of joint production and joint consumption. These are important features of human behavior which are not adequately treated by our analysis.[24] But in this respect our model does no worse (and no better) than the current model of household production (Pollak and Wachter 1975).

Finally, in the empirical part of this paper I have explored only a small fraction of the implications of the model for the allocation of time and consumption patterns of the household. Topics such as the interaction between work at home and substitute market services or the interaction between entry costs (time and money) and time use have only been touched upon. More ambitious endeavors, such as the estimation of the household production function and the value of home output, are still in a preliminary stage. However, given the right data, it is hoped that this paper will facilitate their realization.

---

[24] It may very well be that the observed goods intensity of children can be traced to joint consumption and production. Much of the satisfaction derived from children and much child-care activity involves just having children around while doing other things, such as cooking or watching TV.

References

Becker, Gary S. "A Theory of the Allocation of Time." *Econ. J.* 75 (September 1965): 493–517.

———. "A Theory of Marriage: Part I." *J.P.E.* 81, no. 4 (July/August 1973): 813–46.

Ben-Porath, Yoram. "Economic Analysis of Fertility in Israel: Point and Counterpoint." *J.P.E.* 81, no. 2, pt. 2, suppl. (March/April 1973): S202–S233.

Bloch, F. "The Allocation of Time to Market and Non-Market Work within a Family Unit." Technical Report no. 114. Inst. Math. Studies Soc. Sci., Stanford Univ., November 1973.

Cogan, John. "Labor Supply with Time and Money Costs of Participation." Rand Corp. R-2044-HEW, August 1976.

Fuchs, Victor R. "Recent Trends and Long-Run Prospects for Female Earnings." *A.E.R. Papers and Proceedings* 64 (May 1974): 236–42.

Gronau, Reuben. "The Intrafamily Allocation of Time: The Value of the Housewives' Time." *A.E.R.* 68 (September 1973): 634–51.

———. "The Allocation of Time of Israeli Women." *J.P.E.* 84, no. 4, pt. 2, suppl. (August 1976): S201–S220. (*a*)

———. "Leisure, Home Production and Work—the Theory of the Allocation of Time Revisited." Mimeographed. Nat. Bur. Econ. Res. Working Paper no. 137. Stanford, Calif., May 1976. (*b*)

Heckman, James J. "Effects of Child-Care Programs on Women's Work Effort." *J.P.E.* 82, no. 2, pt. 2, suppl. (March/April 1974): S136–S263.

Mincer, Jacob. "Labor Force Participation of Married Women." In *Aspects of Labor Economics*, edited by H. Gregg Lewis. Universities-National Bureau Conference Series, no. 14. Princeton, N.J.: Princeton Univ. Press, 1962.

Morgan, James N.; Sirageldin, Ismail A.; and Baerwaldt, Nancy. *Productive Americans: A Study of How Individuals Contribute to Economic Progress.* Ann Arbor: Univ. Michigan Press, 1966.

Nordhaus, William D., and Tobin, James. "Is Growth Obsolete?" In *The Measurement of Economic and Social Performance*, edited by M. Moss. Nat. Bur. Econ. Res. Studies in Income and Wealth, no. 38. New York: Columbia Univ. Press, 1973.

Perlman, R. *Labor Theory.* New York: Wiley, 1969.

Pollak, Robert A., and Wachter, Michael L. "The Relevance of the Household Production Function and Its Implications for the Allocation of Time." *J.P.E.* 83, no. 2 (April 1975): 255–77.

Sharir, S. "The Income Leisure Model: A Diagrammatic Extension." *Econ. Rec.* (March 1975): 93–98.

Sirageldin, I. A. *Non-Market Components of National Income.* Ann Arbor: Univ. Michigan Press, 1969.

Walker, K. E., and Gauger, W. H. "The Dollar Value of Household Work." Mimeographed. New York State Coll. Human Ecology, Cornell Univ., 1973.

Willis, Robert J. "A New Approach to the Economic Theory of Fertility Behavior." *J.P.E.* 81, no. 2, pt. 2, suppl. (March/April 1973): S14–S64.

The Principle of Unanimity and Voluntary Consent in Social Choice

Vernon L. Smith
University of Arizona

A discrete version of the author's incentive-compatible Auction Mechanism for public goods is applied to the problem of social choice (voting) among distinct mutually exclusive alternatives. This Auction Election is a bidding mechanism characterized by (1) unanimity, (2) provision for the voluntary compensation of voters harmed by a winning proposition, and (3) incentives for "reasonable" bidding by excluding members of a collective from maximal increase in benefit if they fail to agree on the proposition with largest surplus. Four of five experiments with six voters, bidding privacy, monetary rewards, and cyclical majority rule structure choose the best of three propositions.

Solutions to the problem of specifying incentive-compatible mechanisms for the provision of public goods have been proposed by Thompson (1965), Groves (1969, 1973), Clarke (1971), Drèze and de la Vallée Poussin (1971), Groves and Ledyard (1975), and Smith (in press *a*, in press *b*). Tideman and Tullock (1976) have applied the Clarke-Groves "demand revealing process" to the problem of social choice among discrete alternatives. Several laboratory and field experimental studies of these mechanisms, and of the so-called free-rider problem in public goods, have been conducted by Bohm (1972), Scherr and Babb (1975), the Public Broadcasting Service (reported by Ferejohn and Noll 1976), and Smith (in press *b*). These studies provide strong support for the proposition that practical decentralized mechanisms exist for the provision of public goods.

The sequel will briefly review the incentive-compatible Auction Mechanism (Smith in press *b*) for a public good whose size is variable. A

Research support by the National Science Foundation is gratefully acknowledged. I am indebted to R. Auster and G. Tullock for stimulating my interest in the application of the Auction Mechanism to the social choice problem and to an anonymous referee for helpful, clarifying suggestions.

[*Journal of Political Economy*, 1977, vol. 85, no. 6]
© 1977 by The University of Chicago. All rights reserved

special case of this mechanism consists of a mutually exclusive and exhaustive set of discrete alternatives. This is the familiar problem of social choice in the sense of voting among discrete alternatives to choose the "preferred" social state. The resulting voting mechanism, which we will call the Auction Election,[1] can be interpreted as an implementation of Wicksell's (1896) "principle of unanimity and voluntary consent in taxation." It is a pleasure to acknowledge Wicksell by plagiarizing the title to section 4 of his great paper.[2] In Section V the results of five voting experiments based on the Auction Election will be reported. The experimental design chosen is one for which the majority rule outcome is indeterminate, but for which there is a unanimously preferred alternative achievable with the automatic compensation features of the Auction Election.

I. The Auction Mechanism for Public Goods

Consider a collective composed of I members. Using a partial equilibrium framework,[3] let $V_i(X)$, with measurement normalized so that $V_i(0) = 0$, be the dollar value of a quantity X of the public good to member i and let q be the dollar price of the public good. Although q is assumed constant, the mechanism is easily modified for cases of increasing or decreasing returns (Smith in press *a*). Assume that collective decision must abide by the following rules or institution:

(1) Let each agent i submit a two-tuple (b_i, X_i) consisting of a bid and a proposed quantity with the understanding that his share of cost is $(q - B_i)\bar{X}$, where

$$B_i = \sum_{j \neq i} b_j, \qquad \text{and } \bar{X} = \sum_{k=1}^{I} X_k/I.$$

(2) Each agent has the unqualified right to veto or agree to the cost share $(q - B_i)\bar{X}$ allocated to him by all other agents. He sends the message "agree" by choosing $b_i = q - B_i$, and $X_i = \bar{X}$. He sends the message "veto" if he chooses $b_i \neq q - B_i$ and/or $X_i \neq \bar{X}$.

---

[1] Auster (1976) has proposed the use of "Compensating Elections" to resolve conflicts in social choice. The Auction Election, the Tideman-Tullock "Clarke Tax" Election, and Auster's proposal are identical in purpose although they represent distinct mechanisms.

[2] Also see Buchanan (1959) who, almost alone among modern scholars, has examined and extended Wicksell's ideas on public choice.

[3] More precisely it is assumed that there are no income effects. A general equilibrium two-good (one private, one public) version of the Auction Mechanism using Cobb-Douglas payoff (induced value) functions and a linear production possibility frontier underlies an experimental design currently in research process. This more sophisticated experiment uses the PLATO computer system to program subjects through the decision process. The results of the first several experiments suggest that introducing income effects does not reduce the ability of collectives to reach public good decisions, but the decision outcomes appear to exhibit more sampling variation (across replications with different groups of subjects) than the partial equilibrium experiments reported in Smith (in press *b*).

(3) Group equilibrium prevails if and only if agreement is signaled by every agent i. If such unanimity obtains, $\tilde{X}$ units of the public good are purchased with agent i paying $b_i\tilde{X} = (q - B_i)\tilde{X}$.

Under these rules, i's net benefit is

$$v_i = \begin{cases} V_i(\tilde{X}) - (q - B_i)\tilde{X}, \quad \text{if } b_i = q - B_i, \quad \text{and} \\ \qquad X_i = \tilde{X} = \left(X_i + \sum_{j \neq i} X_j\right)\Big/ I \quad \text{or} \\ \qquad X_i = \left(\dfrac{1}{I - 1}\right) \sum_{j \neq i} X_j, \quad \text{for all } i \\ V_i(0) = 0, \quad \text{otherwise.} \end{cases} \tag{1}$$

Conditions on (b_i, X_i) to maximize v_i, assuming concavity of V_i, are

$$V_i'\left[\left(X_i + \sum_{j \neq i} X_j\right)\Big/ I\right] = q - B_i, \tag{2}$$

$$b_i = q - B_i, \tag{3}$$

$$X_i = \tilde{X}, \quad \text{for all } i. \tag{4}$$

Each i will try to satisfy (2) where the marginal private benefit from his proposal, X_i, is equal to the net private price allocated to him by the "market," while if (3) and (4) are not satisfied, the rules require each i to accept the inferior outcome $V_i(0) = 0$. The rules exclude i unless he agrees to accept the unit cost $q - B_i$, and the group's proposal $\tilde{X}$. This is very similar to the incentive of an economic agent to "meet the market" in a competitive auction to avoid "exclusion." The difference is that under the above rules if any i is excluded, then all are excluded. Whether this difference leads to more extensive strategic gaming and signaling that prevents equilibrium is an empirical question which so far, in the literature cited, has not been an important problem.

Equations (2)-(4) yield the Lindahl equilibrium.[4] Summing (2) over all i, and using (3),

$$\sum_{i=1}^{I} V_i'(X_i) = \sum_{i=1}^{I} V_i'(\tilde{X}) = \sum_{i=1}^{I} b_i = b_i + B_i = q$$

Hence $\tilde{X}$ must be the Lindahl optimal quantity of the public good.

---

[4] These conditions also define one of many local Nash equilibria. As shown in Smith (in press *b*) the Auction Mechanism provides multiple local Nash equilibria among which is the Lindahl equilibrium. This is easy to see from eq. (1). If X^o is the Lindahl equilibrium, then any $X \neq X^o$ yielding $V_i(X) - (q - B_i)X > 0$ for each i is better than nothing and each i has at least some incentive to agree, i.e., to set $b_i = q - B_i$, $X_i = X$. Empirically (Smith, in press *b*) across experiments with different subjects, the final outcome bids are tightly distributed around the Lindahl prices, $V_i'(X^o)$, so that there is a clear tendency for the Lindahl optimal Nash equilibrium to prevail. But in *every* case the Lindahl (and Pareto) optimal *quantity* is chosen. Hence, empirically, the Auction Mechanism allows for a fair degree of variability in the ex post distribution of wealth while preserving Pareto efficiency.

Observe that if X is a public "bad" for any i, then V_i and V_i' are negative, his bid b_i must be negative, and in equilibrium each member harmed by X is compensated. It is in the interest of all harmed agents to agree as soon as their compensation is "adequate" (what this means theoretically is that marginal Lindahl rent is zero).

II. The Auction Election

In social choice group decision produces a common outcome. Thus a referendum legalizing marijuana either passes or fails, and all citizens experience the resulting state. This is just a discrete public good (or bad) to which the Auction Mechanism applies if X is restricted to assume only the values zero or one. Under unanimity i receives net benefit $v_i = V_i(1) - (q - B_i) = V_i(1) - b_i > V_i(0) = 0$. In the problem of pure "political" choice the proposition has a zero resource price. Thus if the proposal is to make daylight saving time official across the United States, then $q = 0$ for that common outcome. Presumably this issue involves only private valuations. Of course q could be negative, for example, legalizing marijuana saves enforcement costs.

Consider first the case of a choice between two alternatives, proposition A or $\bar{A}$ (not A), that is, if A fails to be approved the status quo continues. In the Auction Election ($q = 0$) the social choice problem is solved by allowing each i to submit a bid $b_i \gtreqless 0$ "on" or "for" the proposition. Proposition A wins if $\sum_{i=1}^{I} b_i \geq 0$, otherwise it loses. As in (1), net benefit for voter i is

$$v_i = \begin{cases} V_i(A) - b_i, & \text{if } b_i \geq -B_i, \quad \text{for all } i \\ V_i(\bar{A}), & \text{otherwise.} \end{cases} \tag{5}$$

A transparent theorem is the following: If

$$\sum_{i=1}^{I} V_i(A) > \sum_{i=1}^{I} V_i(\bar{A})$$

(A yields a larger rent to the collective than $\bar{A}$), then there exists a bid b_i^* for each i such that proposition A will win, that is,

$$\sum_{i=1}^{I} b_i^* \geq 0,$$

and no i will be made worse off, that is, $v_i^* = V_i(A) - b_i^* \geq V_i(\bar{A})$. Formally, each voter i will rationally bid no more than his personal valuation of A, $V_i(A)$, net of the *opportunity cost* of A, $V_i(\bar{A})$, that is, $b_i \leq V_i(A) - V_i(\bar{A})$. Hence, if

$$\sum_{i=1}^{I} [V_i(A) - V_i(\bar{A})] > 0,$$

TABLE 1

VOTER VALUATIONS

VOTER	ALTERNATIVE A	ALTERNATIVE $\bar{A}$
1	30	0
2	0	60
3	40	0
Total	70	60

there exists an $\varepsilon_i \geq 0$ for each i such that if we set $b_i^* = V_i(A) - V_i(\bar{A}) - \varepsilon_i \leq V_i(A) - V_i(\bar{A})$, then

$$\sum_{i=1}^{I} b_i^* = \sum_{i=1}^{I} [V_i(A) - V_i(\bar{A}) - \varepsilon_i] \geq 0 \text{ (proposition } A \text{ wins)},$$

and $v_i^* = V_i(A) - b_i^* = V_i(A) - [V_i(A) - V_i(\bar{A}) - \varepsilon_i] = V_i(\bar{A}) + \varepsilon_i \geq V_i(\bar{A})$.[5]

An example, borrowed from Tideman and Tullock (1976), will help to illustrate the process. Table 1 exhibits the valuations of A and $\bar{A}$ for each of three voters. Proposition A might be a proposed change in a land zoning ordinance. Voter 1, now holding valueless property, would enjoy a capital gain of 30. Voter 2, owning land worth 60, would be wiped out. Voter 3, owning land worth nothing, would receive a gain of 40. In an Auction Election for A voter 1 will bid no more than 30, voter 2 no more than -60, and voter 3 no more than 40. Proposition A yields 10 units more collective rent than $\bar{A}$. Voters can submit bids for A yielding an aggregate net benefit up to 10 without causing A to lose. Thus if $\varepsilon_i = 3$ for each i, then

$$\sum_{i=1}^{3} b_i^* = 1,$$

proposition A wins, and each voter is 3 units better off. Each has a competitive incentive to bid enough to ensure passage of A to avoid

---

[5] *Clearly there is an infinite number of imputations of the surplus*

$$\sum_{i=1}^{I} [V_i(A) - V_i(\bar{A})]$$

among I voters, and each imputation represents a possible equilibrium outcome. Any particular equilibrium outcome may depend upon the strategic or "bluffing" behavior of agents, the utility of the outcome relative to the perceived subjective costs of "strategizing," and so on. The situation is analogous to the so-called indeterminacy in the classical bilateral bargaining problem. Indeterminacy simply means that we do not have as yet a sufficiently sharp behavioral theory to account for particular outcomes within the set of Pareto superior points.

"exclusion" from A and thereby being forced to accept a less desirable alternative.

Obviously, proposition A will fail to win if one or more voters is "too greedy." If voter 1 bids less than 20, thus holding out for more than a 10 unit improvement over his present position, he will block passage and get nothing. If voter 2 bids less than -70, hoping for an improvement in excess of 10, he will get no improvement. And so each voter has an incentive to yield sufficiently to ensure that proposition A wins.

III. The "Clarke Tax" Election

Using this example the Auction Election can be compared instructively with the "Clarke Tax" election proposed by Tideman and Tullock (1976). Under their scheme A would also win over $\bar{A}$. Assuming that each voter responds to the Tideman-Tullock incentive to bid his full valuation, then voter 1 pays a tax of $60 - 40 = 20$, the amount necessary to bring the bids for A up to equality with the bids for $\bar{A}$. Similarly, voter 3 pays a tax of 30, but voter 2 pays no tax because his bids do not change the outcome, that is, A would win over $\bar{A}$ if 2 did not bid at all. Hence, voter 1 ends with a benefit net of his tax, 10, voter 3 with a net benefit, 10; but voter 2, who pays no tax and gets zero from proposition A, has had his property (worth 60 under $\bar{A}$) confiscated without due process. As noted by Tideman and Tullock (1976, p. 1149), "It may seem that a person who sustains a large loss when his preference is not followed deserves compensation, but this cannot be given without motivating an excessive statement of differential value. . . . In regard to the uncompensated losses that are produced, the demand revealing process is similar to majority rule." In the Auction Mechanism, agreement requires those who would gain to compensate those who would lose. Both types of voters have motivation to not ask for an excessive share of the rent, that is, to bid sufficiently high to avoid a less favorable outcome. Any mechanism that does not exploit this competitive-like exclusion characteristic (or in some other manner maintains incentive compatibility) and does not require unanimity seems likely to be inefficient or confiscatory or both (as with majority rule). But these criticisms of the "Clarke Tax" should not detract from the innovative contribution of Clarke, Tideman, and Tullock. It should be expected that any of the several new mechanisms of public choice now in only their formative stages will be subjected to a variety of improvements.[6]

---

[6] The Auction, "Clarke Tax," and Auster "Compensating" elections differ primarily in terms of how each would allocate the consumer surplus from an issue among the electorate.

IV. Multiple Choice Auction Elections: An Experimental Design

The Auction Election extends readily to multiple options. Let there be N discrete alternatives, $p = 1, 2, \ldots, N$, with option p having value V_i^p normalized so that the status quo null alternative to the N options has value zero for all i. Each i submits a bid b_i^p for option p. A winning option is one for which the algebraic sum of the bids is nonnegative and no smaller than the algebraic sum of the bids for any other option. Tied options are considered equivalent and a winner selected by an equal probability random device. The expression for net benefit is then[7]

$$v_i = \begin{cases} V_i^p - b_i^p & \text{if } 0 \leq b_i^p + B_i^p \geq b_i^q + B_i^q, \\ & \qquad \text{for all } i \text{ and all } q \neq p, \\ 0, & \text{otherwise,} \end{cases} \tag{6}$$

where

$$B_i^p = \sum_{j \neq i} b_j^p, \qquad B_i^q = \sum_{j \neq i} b_j^q.$$

If there is a winning option p it has the property

$$0 \leq \sum_{i=1}^{I} b_i^p \geq \sum_{i=1}^{I} b_i^q$$

for all $q \neq p$. If there is no winning option, that is, $b_i^r + B_i^r < 0$ for all $r = 1, 2, \ldots, N$ the status quo is maintained.

The example in table 2 applying to six voters and three options provides the subject valuations in dollars used in the experiments to be discussed in the next section. The entries in table 2 are measured relative to the status quo, or the null proposition (not shown). Thus if proposition 3 wins, voter 1's wealth state is \$20 lower than if none of the three propositions wins. Propositions 2 and 3 are both highly attractive with aggregate rents of \$105 and \$45, respectively, but the Auction Election can be expected to choose 2. Under a majority rule decision process the outcome is intransitive. Proposition 2 beats 1 by a vote of 4 to 2, proposition 3 beats 2 by 4 to 2, and 1 beats 3 by 4 to 2. As is well known and typical of majority rule, (1) it leads to an inefficient outcome unless the "right" agenda is followed, and there is no way (to my knowledge) that an appropriate agenda can be selected on the basis of objective voting data, and (2) even if the optimal choice is selected by majority rule it produces an involuntary redistribution of wealth. The Auction Election redistributes relative wealth voluntarily.

---

[7] Note that an Auction Election with N options is just the Auction Mechanism with N discrete, mutually exclusive alternatives, i.e., N different public goods $X = 1, 2, \ldots, N$, one and only one of which can be chosen.

TABLE 2

VOTER VALUATIONS ($)

PROPOSITION	VOTER 1	2	3	4	5	6	TOTAL
1	5	−30	−30	25	25	0	−5
2	60	5	5	−10	−10	55	105
3	−20	45	45	0	0	−25	45

V. Experiments and Results

Thirty subjects participated in five experimental sessions, each consisting of six-member voter collectives using the induced value payoffs[8] in dollars shown in table 2. The instructions printed in the Appendix were distributed and read aloud to the six subjects in each session who had been recruited from large sections of business and economics courses at Arizona State and the University of Arizona. No subject participated in more than one session. Each subject also received a copy of a recording form, a sample of which (for voter 1 in table 2) is included in the Appendix. The three propositions, as defined by the vector of subject valuations, were reordered (i.e., renumbered 1, 2, 3) randomly for each experiment. In each experiment the subjects were assigned randomly to the six valuation conditions and seated so that the privacy of such information could be maintained. This right of privacy included a provision for separately paying each subject his cash earnings after each experiment was completed. It was then each subject's choice whether such information was to be revealed to any other person.

In each experiment a maximum number of trials was specified. The first trial was a "practice trial" that did not count in the determination of a winning proposition. Experiments 1 and 2 (see fig. 1) consisted of a maximum of 10 trials, while experiments 3, 4, and 5 allowed up to six trials. The trial maximum was reduced after the first two experiments as it became evident that a very few trials were sufficient to allow collectives to reach a decision. It seems possible that more trials may even make group agreement less likely. Only one of the five experiments (experiment 2) failed to produce a winning proposition, and it was a 10-trial session.

Figure 1 exhibits the trial sequence of individual subject bids for the superior proposal (proposition 2, table 2) in each experiment. The subject identification numbers correspond to the voter numbers, with associated

---

[8] Subjects were paid $V_i^p - b_i^p + \$2$ if proposition p won, $V_i^o = \$2$ if no proposition won, where the induced values V_i^p were as given in table 2. A total of $464 was paid to the 30 subjects under these rules.

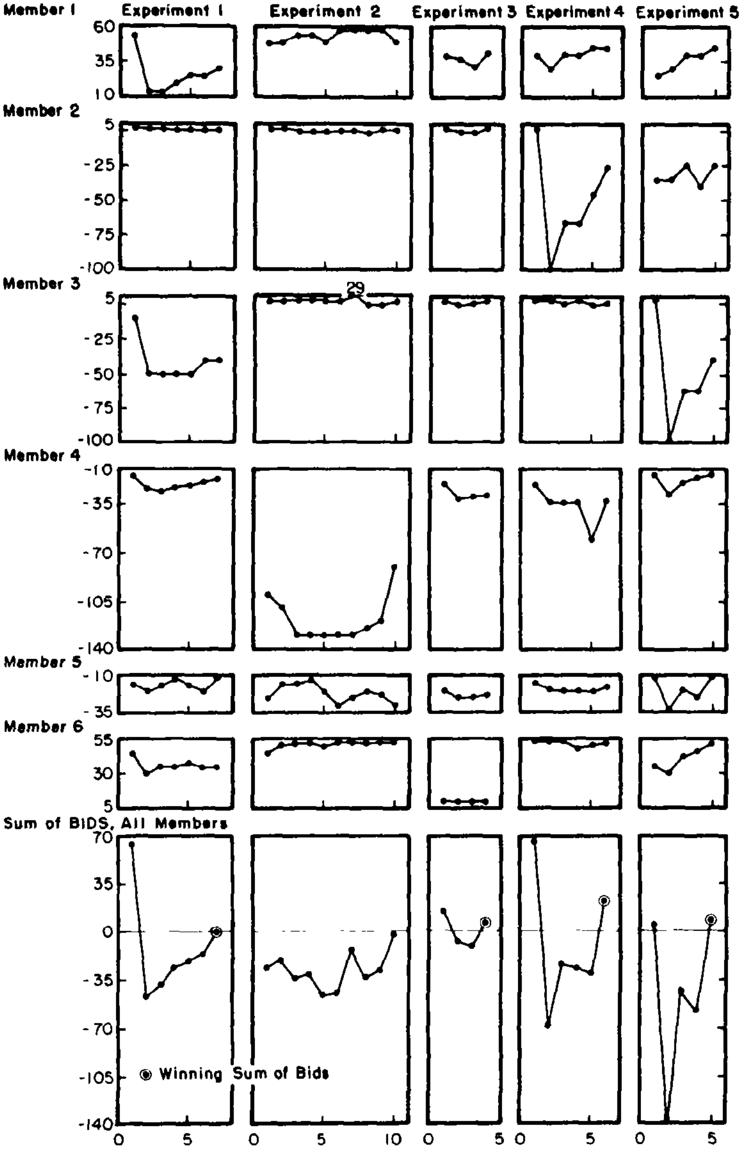

Member 1
Experiment 1
Experiment 2
Experiment 3
Experiment 4
Experiment 5
60
35
10
Member 2
5
-25
-50
-75
-100
Member 3
29
5
-25
-50
-75
-100
Member 4
-10
-35
-70
-105
-140
Member 5
-10
-35
Member 6
55
30
5
Sum of BIDS, All Members
70
35
0
-35
-70
-105
-140
◉ Winning Sum of Bids
0
5
0
5
10
0
5
0
5
0
5

FIG. 1

proposition values, shown in table 2. The panel at the bottom of figure 1 plots the bid sum for each trial.

The practice trial, besides providing familiarity with the procedure, was of considerable importance in yielding information, however imperfect, on the potential surplus available among the three proposals. In every experiment the best proposition received the largest bid sum on the first trial. This first-trial bid sum was positive in all sessions except experiment 2 and (especially in experiments 1 and 4) revealed to members of the collective that the total surplus available from the best proposition was substantial. Consequently, on the second trial the bids tended to fall sharply as subjects generally attempted to "game" for an increased share of the surplus (23 of the 30 subjects reduced their bids on trial 2). The question or hypothesis at issue, as in previous public good experiments (Smith, in press *b*) is *not* whether members will "game" for a personal advantage in the Auction Election, but whether the mechanism is sufficiently effective to guide the collective to the optimal choice. Obviously some, perhaps many, members will "game"; and obviously, in figure 1, they do "game." But the mechanism also confronts each member of the collective with the discipline of exclusion if each does not, by the final trial, at least match the "market price," $-B_i$, presented to him by the other members. Four of the five experimental collectives reached agreement on the optimal proposition. In experiment 2 the group failed by $2,

$$\sum_{i=1}^{6} b_k = -2,$$

on trial 10 to elect the best proposition. If we take these results literally, the sample evidence suggests that an Auction Election choice among three propositions (four, counting the null proposition or status quo) with parameters as defined in table 2 would select the optimal proposition in 80 percent of the elections. (And remember, under majority rule *none* of these propositions can prevail by pairwise voting.) The robustness of these results can only be established with more experiments. Experiment 2 was very close to producing agreement in spite of unusually aggressive, and ineffective, play by subject 4. This subject expressed puzzled disappointment with the outcome, suggesting that failure experiences might mitigate such aggressive bidding in subsequent elections. Experience may be an important treatment variable and deserves systematic study. However, the failure of experiment 2 may be due to the fact that subject 3's bid (29) on trial 7 exceeded his valuation (5), producing unrealistic profit expectations in the other subjects.

In previous public good experiments (Smith, in press *b*) using the Auction Mechanism it was hypothesized that equilibrium bids across

TABLE 3

Trial	$\hat{\beta}$	$t_{\beta-1}$	$\hat{\beta}_o$	$t_{\beta o}$	R^2	SE	$F(1,28)$
1	0.92	−0.678	−12.0	−3.01	.674	18.71	60.87
$t^* - 1$	1.105	−0.767	−24.18	−4.59	.630	24.72	50.38
t^*	0.998	−0.019	−16.33	−4.39	.737	17.46	82.29

individual subjects would tend to the theoretical Lindahl prices for those subjects. Formally this took the form of testing the research hypothesis $\beta_o = 0$ and $\beta = 1$ in the regression equation

$$y_{ij} = \beta_o + \beta x_{ij} + \eta_{ij}, \tag{7}$$

where y_{ij} is the observed final equilibrium bid of subject j in experiment i and x_{ij} is the (experimentally controlled) Lindahl price for subject j in experiment i. In the present Auction Election experiments there are no Lindahl prices where marginal Lindahl rents are zero because the public outcome of an election is discrete. However, the individual valuations x_{ij} of subject i in experiment j for the best proposition do represent Lindahl price upper bounds on subject equilibrium bids. In the Auction Election these valuations should explain much of the variation in the final outcome bids. In the above regression for the Auction Election the research hypothesis becomes $\beta_o < 0$, $\beta = 1$. If $\beta = 1$ the estimate $-\hat{\beta}_o$ is a measure of the equilibrium surplus ($x_{ij} - y_{ij} = -\beta_o$) obtained by the average subject. In the Auction Election the hypothesis that $\beta = 1$ follows from the expectation that the surplus obtained by a subject on a winning proposal will not depend on his valuation, that is, those able to pay z dollars more will on the average pay z dollars more.

The regression results using the data shown in figure 1 for the first, penultimate, and final (t^*) trials are shown in table 3. For each of these trials $\hat{\beta}$ is very close to unity with very low t-values, particularly on the final trial. The estimates of $\hat{\beta}_o$ indicate that the average voter's initial demand for net surplus, \$12, is below his final earnings, \$16. On the penultimate trial the extent of "bluffing" is indicated by an average demand for \$24 of surplus. The R^2 values show that subject valuations account for 63 percent or more of the variation in subject bids. However, the final trial R^2 of .74 is below the final trial R^2 in three different sets of public good experiments (Smith 1976*b*), in which the Lindahl prices accounted for 82 percent, 97 percent, and 99 percent of the variation in subject bids. A reasonable conjecture might be that the mechanically simpler Auction Election invites more strategic bluffing and signaling by members of the collective.

VI. Conjecture, Impossibility Theory, and the Evidence

In 1896, Wicksell conjectured that if the utility of a public outcome exceeded its cost then it was theoretically possible to find a distribution of costs so that all members of the collective would approve the outcome unanimously. In 1955, Samuelson first conjectured the "fatal inability" of any decentralized mechanism to determine optimal public choices, a view that was later (Samuelson 1969) reasserted even more emphatically. Neither Wicksell nor Samuelson actually specified decentralized mechanisms which could provide theoretical foundations for their respective conjectures. The Auction Election solves the Arrow (1963) problem in somewhat the same sense that a competitive private goods economy solves the problem of resource allocation: the competitive economy provides a Pareto optimal outcome for a given distribution of primary resources and has nothing to say about the optimal distribution of such resources. Actually, the mechanism does a little more than this by providing for some voluntary redistribution of wealth whenever a winning proposition is produced.

Hurwicz (1972), for pure exchange private goods economies, and Ledyard and Roberts (1974), for economies with public goods, have proved that it is impossible to find a mechanism that provides individually rational Pareto optima and which is simultaneously individually incentive compatible. Obviously, the above mechanisms do not contradict these theorems which postulate a much broader range of strategic behavior by economic agents than Nash equilibrium competitive behavior. But why do real people making real decisions for real money, in the experiments reported here and in the cited experimental public good literature, behave predominantly in accordance with the competitive postulate? Why do they not exhibit the more "sophisticated," "strategic" behavior postulated by Hurwicz and Ledyard-Roberts? I think it is because there are significant direct (and indirect opportunity) costs of thinking, calculating, and signaling which make strategizing uneconomical. In the Auction Election the opportunity cost of failing to reach agreement is the loss of the more valuable best proposition. Strategizing not only consumes costly time and thought, it increases the risk of group disagreement. Disagreement means exclusion from better wealth states on the sobering principle that nobody gets more if there is not more available to get.

Can all the hard scientific evidence from experiments to date be dismissed as irrelevant, too simplistic, or based on particular experimental parameters? I think not, for the reason that such results are consistent with a great deal of field evidence. Thousands of buildings have been purchased by religious organizations, clubs, art associations, and private societies through member voluntary contributions having the exclusionary characteristic of the Auction Mechanism and the Auction Election: if contributions do not cover the cost of the proposed project, the project dies. Numerous universities meet specific capital needs from contributions and gifts by alumni and associates. We tend to explain these phenomena, while holding on to our enshrined belief in the impossibility of decentralized public good decision, by attributing them to atypical "altruism." The theory discussed here suggests that such behavior is entirely consistent with self-interest motivation since the decision procedure is the same as that of the Auction Mechanism for public goods.

But it is not my intention to argue that either the present state of evidence or the present state of theory is satisfactory. There is no theory, only conjecture, to support the assertion above that competitive behavior may be a rational response when signaling, thinking, and calculating are costly. Nor can we be confident of the empirical results until new experiments are conducted and old experimental results replicated with different subjects and different experimenters. In future experiments I intend to explore the effect of larger collectives, experience, and balanced budget versions of the Auction Election in which aggregate overbids are rebated in proportion to individual bids.[9]

Appendix

Instructions

This is an experiment in the economics of group decision making. The instructions are simple, and if you follow them carefully and make good decisions you may earn a considerable amount of money which will be paid to you in cash at the end of the experiment. Various research foundations have provided funds for this research.

You are a member of a group that must decide which one of 3 proposals is to be selected. The group will decide by bidding which proposal will be chosen. The value to you of each alternative proposal is shown in column (1) (under each

---

[9] In earlier experiments (Smith, in press *b*) it was found that the stopping requirement under which each member must *exactly* match his share of cost was needlessly demanding on group agreement. A better rule is to require $b_i \geq q - B_i$ and to distribute any overbid as a rebate to each member in proportion to his bid. Letting

$$B = \sum_{k=1}^{I} b_k = b_i + B_i > q,$$

each i is assigned a new lower bid

$$b_i' = b_i - \frac{b_i(B - q)}{B}.$$

Then

$$B' = \sum_{i=1}^{I} b_i' = \sum_{i=1}^{I} \left[b_i - \frac{b_i(B - q)}{B} \right] = q.$$

The budget is balanced with the new bids (b_i'), and since $b_i' < b_i$ every agent is better off. In any case any agent can still veto the arrangement and reject the new calculated bid. This procedure makes group agreement mechanically easier, i.e., compliance cost is lower for the group stopping rule.

TRIAL 1

PROPOSITION	(1) MY VALUE	(2) MY BID	(3) MY NET VALUE
1	5		
2	60		
3	-20		
4			
5			

TRIAL 2

PROPOSITION	(1) MY VALUE	(2) MY BID	(3) MY NET VALUE
1	5		
2	60		
3	-20		
4			
5			

TRIAL 3

PROPOSITION	(1) MY VALUE	(2) MY BID	(3) MY NET VALUE
1	5		
2	60		
3	-20		
4			
5			

TRIAL 4

PROPOSITION	(1) MY VALUE	(2) MY BID	(3) MY NET VALUE
1	5		
2	60		
3	-20		
4			
5			

TRIAL 5

PROPOSITION	(1) MY VALUE	(2) MY BID	(3) MY NET VALUE
1	5		
2	60		
3	-20		
4			
5			

TRIAL 6

PROPOSITION	(1) MY VALUE	(2) MY BID	(3) MY NET VALUE
1	5		
2	60		
3	-20		
4			
5			

FIG. XI

trial), on your record sheet. Net value, which will be paid to you in cash on the final decision, is computed by subtracting your bid from your value for the winning proposal.

The decision process consists of a series of trial bids as follows (refer to your record sheet): On each trial each member privately selects and writes a bid in column (2), expressed in whole dollars, for each of the proposals. For example, 20, 0, -15. A negative bid means that you are asking for compensation from the other members. *A bid should not exceed your proposal value if you want to make money. On each trial, I will go to each member and record his bids, compute the algebraic sum of all member bids for each proposal, and post these sums on the blackboard.* The proposal with the largest positive bid sum will be the winner. On each trial compute your potential net value, column (1) minus column (2), for each proposal and write it in column (3). If and when a proposal wins, draw a circle around your net value for the winning proposal. You will have a maximum of 6 trials to determine the winning proposal. The first trial will be for practice to familiarize you with the procedure and will not count in determining a winning proposition.

If a proposal wins we will stop on that trial, and you will be paid $2 plus your net value. Otherwise, you will be paid $2. The proposition values on your record sheets are not the same for all members. They represent your own private information and are not to be revealed to any other member. Feel free to earn as much cash as you can. Do not speak to any other participant.

References

Arrow, K. *Social Choice and Individual Values*. 2d ed. New York: Wiley, 1963.

Auster, R. "Renting the Streets." *Proceedings of the Conference on American Re-Evolution*. Tucson: Dept. Econ., Univ. Arizona, in press.

Bohm, P. "Estimating Demand for Public Goods: An Experiment." *European Econ. Rev.* 3 (1972): 111–30.

Buchanan, J. "Positive Economics, Welfare Economics, and Political Economy." *J. Law and Econ.* 20 (October 1959): 124–38.

Clarke, E. "Multipart Pricing of Public Goods." *Public Choice* 11 (Fall 1971): 17–33.

Drèze, J., and de la Vallée Poussin, D. "A Tâtonnement Process for Public Goods." *Rev. Econ. Studies* 38 (April 1971): 133–50.

Ferejohn, J., and Noll, R. "An Experimental Market for Public Goods: The PBS Station Program Cooperative." *A.E.R. Papers and Proceedings* (May 1976): 267–73.

Groves, T. "The Allocation of Resources under Uncertainty. The Informational and Intensive Roles of Prices and Demands in a Team." Technical Report no. 1, Univ. California Berkeley, Center Res. Management Sci., August 1969.

———. "Incentives in Teams." *Econometrica* 41 (July 1973): 617–33.

Groves, T., and Ledyard, J. "Optimal Allocation of Public Goods: A Solution to the 'Free-Rider Problem.'" Discussion Paper no. 144, Northwestern Univ., Center Math. Studies Econ. and Management Sci., September 1975.

Hurwicz, L. "On Informationally Decentralized Systems." In *Decision and Organization*, edited by R. Radner and B. McGuire. Amsterdam: North-Holland, 1972.

Ledyard, J., and Roberts, J. "On the Incentive Problem with Public Goods." Discussion Paper no. 116, Northwestern Univ., Center Math. Studies Econ. and Management Sci., 1974.

Samuelson, P. "Diagrammatic Exposition of a Theory of Public Expenditure." *Rev. Econ. Statis.* 37 (November 1955): 350–56.

———. "Pure Theory of Public Expenditure and Taxation." In *Public Economics*, edited by J. Margolis and H. Guitton. New York: St. Martins, 1969.

Scherr, B., and Babb, E. "Pricing Public Goods: An Experiment with Two Proposed Pricing Systems." *Public Choice* (Fall 1975): pp. 35–48.

Smith, V. "Mechanisms for the Optimal Provision of Public Goods." *Proceedings of the Conference on American Re-Evolution*. Tucson: Dept. Econ., Univ. Arizona, in press. *(a)*

———. "Incentive Compatible Experimental Processes for the Provision of Public Goods." NBER Conference on Decentralization, April 23–25, 1976. Forthcoming in *Research in Experimental Economics*, edited by V. Smith. Greenwich, Conn.: JAI Press, in press. *(b)*

Thompson, E. "A Pareto Optimal Group Decision Process." *Papers in Non-Market Decision-Making* 1 (1965). 133–40.

Tideman, T., and Tullock, G. "A New and Superior Process for Making Social Choices." *J.P.E.* 84 (December 1976): 1145–59.

Wicksell, K. "A New Principle of Just Taxation." Translated by J. Buchanan. In *Classics in the Theory of Public Finance*, edited by R. Musgrave and A. Peacock. New York. St. Martins, 1967.

An Economic Analysis of Marital Instability

Gary S. Becker and Elisabeth M. Landes
University of Chicago and National Bureau of Economic Research

Robert T. Michael
Stanford University and National Bureau of Economic Research

This paper focuses on the causes of marital instability. Section I develops a theoretical analysis of marital dissolution, incorporating uncertainty about outcomes of marital decisions into a framework of utility maximization and the marriage market. Section II explores implications of the theoretical analysis with cross-sectional data, primarily the 1967 Survey of Economic Opportunity and the Terman sample. The relevance of both the theoretical and empirical analyses in explaining the recent acceleration in divorce rates is also discussed.

At the beginning of this century, separation and divorce were unimportant sources of marital dissolution[1] compared to death from childbirth, contagious diseases, and other causes. Couples marrying could expect to remain together until death. The substantial decline in death rates during this century, combined with a steady growth in separations and divorces that sharply accelerated during the last 10 years, has radically altered these expectations. Today, a typical couple has only a small probability of being separated by death during their first 15 years of

---

This study has been partially supported by grants to NBER from the National Institute of Child Health and Human Development, DHEW; The Rockefeller Foundation; the National Science Foundation for the study of law and economics; and by NBER. The authors wish to thank Michael Grossman, Michael Keeley, William Landes, Sam Peltzman, George Stigler, participants in workshops at the University of Chicago, Columbia University, NBER-West, and University of Rochester, and a referee for comments, and Barbara Andrews, Kyle Johnson, Gale Mosteller and Richard Wong for valuable research assistance. This report has not undergone the review accorded official NBER publications.

[1] Throughout this study we use the terms "divorce" and "dissolution" interchangeably, and we do not distinguish in the theoretical section among separation, annulment, and divorce.

[*Journal of Political Economy*, 1977, vol. 85, no. 6]
© 1977 by The University of Chicago. All rights reserved

marriage, but perhaps 10 times as high a probability of being separated by divorce.

This dramatic change in the incidence of voluntary dissolutions has major implications for many kinds of family behavior. Couples are reluctant to invest in skills or commodities "specific" to their marriage if they anticipate dissolution: having children and working exclusively in the nonmarket sector are two such marriage-related activities. That is, the rise in women's labor force participation rates and the fall in fertility rates in the past 2 decades have partly been caused by, as well as being causes of, the rise in marital instability.

Although many effects of marital dissolution are discussed, this paper focuses on the causes of dissolution. Why are divorces more common among the poor, blacks, geniuses, and the retarded, or among couples marrying young, or couples in racially or religiously mixed marriages? Do the causes of cross-sectional differences in divorce also explain the growth in the divorce rate over time, including its recent acceleration?

We believe that these causes can be discovered by extending the analysis of marriage developed by Becker (1974). He assumes that persons marry when the utility expected from marriage exceeds the utility expected from remaining single. It is natural to assume further that couples separate when the utility expected from remaining married falls below the utility expected from divorcing and possibly remarrying. One way to reconcile the relatively high utility expected from marriage at the time of marriage and the relatively low utility expected at the time of dissolution is to introduce uncertainty and deviations between expected and realized utilities. That is, persons separating presumably had less favorable outcomes from their marriage than they expected when marrying.

Section I develops a theoretical analysis of marital dissolution that incorporates uncertainty about outcomes of marital decisions into the framework of utility maximization and the marriage market. This analysis has implications about the effects of income, age at marriage, fecundity impairments, number of children, duration of marriage, welfare payments, and many other variables on the likelihood of marital dissolution. Section II tests these implications with cross-sectional data, primarily the 1967 Survey of Economic Opportunity and the Terman sample of geniuses. Evidence from many other studies and from time series is also discussed. For the most part, the evidence confirms the theoretical predictions.

The analysis presented here is also applicable to other contracts of indefinite duration, where the parties involved have the option of termination, perhaps with a penalty. Examples include explicit contracts between business partners and implicit "contracts" binding together employees and employers, customers and suppliers, or friends. The relation, for example, of employee turnover to duration of employment, specific investments, marital status, and other variables is illuminated by the analysis in this paper.

I. Theoretical Analysis

1. *Basic Framework*

Households are assumed to use nonmarket time and market goods to produce nonmarketable commodities. Each person maximizes the utility from the commodities that he or she expects to consume over his lifetime. With risk-neutrality, this criterion simplifies to the maximization of expected full wealth—the present value of the stream of commodities consumed. Full wealth does not equal money wealth alone but also takes account of the productivity of nonmarket time.

By assumption, each marital "strategy" produces a known amount of full wealth, and the opportunity set equals the set of full wealths produced by all conceivable marital strategies. The individual ranks all strategies by their full wealth and chooses the highest. Even with certainty, a strategy with marriage, then dissolution, and eventually remarriage might be preferred to all other strategies and would be anticipated at the time of first marriage. Dissolution would be a response perhaps to the growing up of children, or to diminishing utility from living with the same person, and would be a fully anticipated part of the variation in marital status over the life cycle.

It is commonplace that uncertainty pervades all decisions, and perhaps nowhere has this been more fully appreciated than in discussions about marriage.[2] Even after prolonged dating, newly married persons face tremendous uncertainty about their own or their mate's needs, their capacity to get along with each other, their fecundity and other aspects of having and raising children, and so on almost indefinitely. Uncertainty introduces a whole new dimension into the analysis because dissolution no longer need be fully anticipated but can result from unexpected events.

The optimal marital decision at any moment would be the one that maximized the expected value of full wealth over the remainder of life, given the realizations up to that moment. The optimal strategy would be the set of all these optimal decisions and would in general include divorce at different stages in the life cycle, sometimes contingent on the realization of unfavorable outcomes, and sometimes consistent with the realization of expected outcomes. With divorce viewed in a stochastic framework, it is natural to consider the probability of divorce as a function of two factors: the expected gain from marriage and the distribution of a variable describing unexpected outcomes. The probability of divorce is smaller the greater the expected gain from marriage, and the smaller the variance of the distribution of unanticipated gains from marriage.

---

[2] "Marriage is the only adventure open to the timid" (Voltaire); "marriage be a lottery in which there are a wondrous many blanks . . ." (Vanburgh); "marry in haste, and repent at leisure" (Cabell). (These references are taken from Evans [1968].)

We suggest that the majority of divorces results from uncertainty and unfavorable outcomes and, therefore, would not occur in a world where outcomes could be anticipated. Indirect evidence supporting this view is that most dissolutions occur early in marriage, not after many years when children have grown or couples have tired of each other. In fact, the median duration to divorce has been about 7 years, and three-quarters of all divorces take place before the fifteenth anniversary of marriage.[3] Since there are sizable emotional and financial costs of divorcing, people would presumably prefer to remain single rather than enter a marriage that is expected to dissolve within a few years.

Up to this point we have discussed one spouse's decision about divorce as if the other spouse had no say in the matter. If the two spouses concur in judging their own expected full wealth to be greater either by remaining married or by divorcing, there would be no disagreement about whether or not to divorce. But what if these judgments differ? If all compensations between spouses were feasible and costless, a couple would separate if, and only if, their combined wealth from remaining married were expected to be less than their combined wealth when separated. For if one spouse expected greater separated wealth while their expected married-wealth was greater than their combined expected separated-wealth, the other spouse would be able to compensate the first to remain married. Likewise, if one spouse expected less separated wealth while their combined separated-wealth was greater than their married-wealth, he or she could be compensated to separate (if consent were required) because the one spouse's gain would exceed the other's loss. Indeed, compensation of a spouse to induce acquiescence is an excellent illustration of the "Coase Theorem" that the allocation of property rights or legal liability does not influence resource allocation when the parties involved can bargain with each other at little cost

The conclusion that a couple dissolves their marriage if, and only if, their combined wealth when dissolved exceeds their combined married-wealth is a direct extension of the conclusion (see Becker 1974) that single persons marry if, and only if, their combined married-wealth exceeds their combined single-wealth. Both conclusions assume that the division of wealth between mates is flexible, which contrasts sharply with the assumption implicit in many discussions, namely, that the division of married-wealth is rigidly determined by custom, "family" goods, and the like. Asset transfers and alimony payments after dissolution introduce more flexibility into the division than may appear from the importance of "family" goods, in the same way that asset transfers prior to marriage—

---

[3] During the 1950s and 1960s the median duration of marriage prior to divorce ranged between 5.8 and 7.5 years in the United States (Plateris 1973*b*, p. 39; and Plateris 1973*a*, p. 49).

such as dowries and bride prices—introduce more flexibility into marital divisions.

If the division of wealth between spouses is sufficiently flexible, it would not be meaningful to say that one mate "walked out" on or was "abandoned" by the other. This is obviously not a useful distinction when each gains from divorce, but it is also true, if less obviously, when divorce is available at the option of either mate. Suppose that one mate gained 100 units and his spouse lost 60 units of real income from a divorce, relative to their *predivorce division of outputs*. Relative to that division, when divorce occurs one might say she was "abandoned" and he "walked out." He would be willing to stay, however, if the division within marriage were changed in his favor by at least 100 units, but with that division she would "walk out" and he would be "abandoned" because she would gain more than 40 units from a divorce, and he would lose. Whether one mate "walks out" or is "abandoned" is ambiguous, therefore, and depends critically on the marital division that is used as a yardstick.

The same argument applies to the distinction between "quits" and "layoffs" in discussions of the turnover of employees. If the combined wealth of a firm and employee were decreased by a separation, there would exist a transfer (i.e., a wage payment) from the firm to the employee (or vice versa) that would induce them to stay together. Of course, even if their combined wealth were increased by separation, the firm would want to keep him and he would want to leave *at some wage*. However, at a sufficiently higher wage, the firm would want him to leave and he would want to stay. Although wage "rigidity" may prevent fluid divisions between firms and employees, the rigidity in labor (as well as marriage) markets has been greatly exaggerated, and combined maximization is probably also the appropriate model in labor markets.[4]

2. *Dissolution and Expected Gain from Marriage*

The probability of divorce is greater the smaller the expected gain from marriage, provided unexpected gains are not strongly negatively correlated with the expected gain. Becker (1974) provides an extensive analysis of optimal marital sorting that explains the predominance of positive assortative mating with respect to personal characteristics such as education, height, intelligence, age, property income, physical attractiveness, etc. The explanation applies to all traits which are not good substitutes in the production of commodity income, while negative assortative mating would be optimal for substitutes, such as wage earning power. Becker further shows that, where positive assortative mating is optimal, persons with higher-valued characteristics gain more from marriage (compared to being single). So couples with, say, more property income or education would be expected to have greater gains from marriage and consequently lower probabilities of divorce.

---

[4] Instead of basing the distinction between quits and layoffs on rigidity in the wage or marital division, a more promising approach relies on the *cause* of a job or marital separation. A quit could be said to result from an improvement in opportunities elsewhere and a layoff from a (usually unexpected) worsening in opportunities in this job or marriage. This way of distinguishing quits from layoffs has many implications, among them that persons quitting have shorter spells of unemployment (or duration of time to remarriage) than persons laid off and improve their circumstances more in their new jobs (or marriages). These implications have received some empirical support (see Kuratani 1973; Bartel 1975; Hashimoto 1975; Martin 1977). We conjecture that most marital separations, as most job layoffs (see Feldstein 1976), are temporary; i.e., separated mates are more likely to return to each other than to divorce.

Becker's analysis of optimal sorting assumes that the traits and productive capacities of persons are fixed. However, they are affected by the marital sorting itself. For example, a person will tend to specialize in acquiring skills that raise market productivity compared to nonmarket productivity if he spends more time in the market sector after marriage as a result of substitution of spouse's time in the nonmarket sector. Conversely, he will specialize more in acquiring nonmarket skills if he spends more time in the nonmarket sector after marriage.

Therefore, the gain from marriage compared to being single depends in part on the extent to which investments in skills are oriented to the division of labor within marriage. The effect of specialized investments on the incentive to become and to stay married can explain why women have traditionally married earlier: their investments have been more closely geared to child rearing, household management, and other domestic activities that are much less useful to single persons.[5] As another example, consider men with relatively high earnings potential. In the optimal sorting, they marry women with relatively low earnings potential, greater physical attractiveness, and superior other nonmarket characteristics. Therefore, men with relatively high earnings potential gain more from marriage than men with relatively low earnings potential not only because of the higher level of their income but also because of greater gains from specialization within marriage, since their mates have a comparative advantage in specializing in nonmarket investments.

As a final example, consider the relation between level of schooling and the gain from marriage. For persons with relatively high levels of schooling, the effect of specialized investments on the gain from marriage at least partly offsets the effect of optimal sorting. On the one hand, marriages between highly educated individuals have greater gains because of the spouses' high levels of market and nonmarket skills. On the other hand, they have lower gains because they typically involve less specialization between spouses, since more educated women participate more in the labor force. Consequently, there is no clear theoretical prediction about the net effect of schooling level on the gain from marriage.

---

[5] Accordingly, it is not surprising that the sex differential in age at first marriage has greatly declined during the last 20 years (see Plateris 1973*b*, p. 55); the investments of women have become much less specialized to married life as they have reduced their childbearing and increasingly entered the labor force.

3. *Dissolution and Search*

In this section we discuss the sorting of persons and the stability of marriages when there is limited information about the traits of potential mates, and when remarriage is not possible (remarriage is introduced in Section I.5). It is often difficult—that is, expensive—in actual marriage markets to find a satisfactory mate. For example, persons with rare traits, such as an IQ over 150, $1 million, a height in excess of 6 feet 6 inches, or being a Moslem in South Dakota, usually have to spend considerable resources "searching" for mates with similar traits because most persons encountered have more typical traits. Anticipating these difficulties, persons with rare traits may compromise and settle for mates with less similar traits; that is, they may give up the gains from an "optimal" mate in order to reduce their expenditures of time and money on search. The costs of finding a satisfactory mate are important in understanding marital dissolutions because they can affect the expected gain from marriage.

Imperfect information that results from the cost of finding a mate cannot increase the gain from marriage above the "optimal" (i.e., the gain with perfect information) for any couple and will reduce the gain for most couples. Since the total gain from marriage over all marriages is maximized in the "optimal" sorting, persons not matched in this sorting could not increase their gain by marrying each other. Consequently, most couples will gain less in all other sortings, and some couples may gain the same amount. The actual and "optimal" sortings differ because the cost of finding a mate induces at least some couples to accept a lower gain from marriage than they would receive in the "optimal" sorting. The larger the marital search costs, the smaller the acceptable gain, and the larger the deviations from the "optimal" sorting. Although all *couples* gain less (or at least do not gain more) than in the "optimal" sorting, some *persons* with relatively low search costs may gain more because they can capitalize on the greater search costs of others to make advantageous marriages.

The process of searching for a mate can be formalized along the lines developed in the extensive recent literature on search.[6] Each person spends resources selecting a drawing from a frequency distribution of potential mates; each drawing gives the wealth that can be expected from that match. This frequency distribution is determined by the search costs of all persons in the marriage market. If search costs were zero for everyone, this distribution would reduce to a single point—the person's wealth in the optimal sorting.

---

[6] Search theory was first applied to the marriage market by Keeley (1974).

After each drawing the individual must decide whether to accept that match or to continue searching for a better one. The cost of continuing to search for a better match is the sum of search costs and any income forgone by remaining single rather than marrying an available match. The expected benefit from continuing to search equals the product of the probability of finding a preferable mate times the expected increase in wealth from finding a preferable mate $(\bar{W}_{mf} - W^a_m)$, where $\bar{W}_{mf}$ is the expected wealth from a better match and W^a_m is the expected wealth in the best available marital status (i.e., single or married to the best available potential spouse). The individual is indifferent to accepting the available offer when the cost and expected benefit from additional search are equal. Actual marital offers differ even among persons with the same search costs and the same frequency distribution of offers since some will be "lucky" and some will be "unlucky." The latter will have, after the marital search process ends, lower expected gains from marriage and thus higher probabilities of divorce.

The search process can also be usefully described in terms of the set of acceptable traits. If search costs, wealth, and number of persons varied continuously as a function of traits, the acceptable traits would form a closed continuous set around the "optimal" trait (i.e., the trait of one's mate in a world with perfect information) (see Wessels 1976). The upper and lower bounds of this set are depicted as A^u and A_l in figure 1. The expected wealth from a match with either trait A_l or trait A^u must equal the value of additional search when these matches are available, and these matches must both provide the same wealth, W^a_{mf}. The offers available from matches to persons with traits anywhere to the left of A_l or to the right of A^u must be less than the value of additional search when faced with these matches (otherwise these traits would be in the acceptable set), so continuity implies equality between the offers and the value of search at the boundaries of the acceptable set. These boundaries are determined not only by one's own search costs, but also by the costs of everyone else in the marriage market, the distribution of traits in this market, and household production functions.

Since offers must exceed the value of search in the interior of the acceptable set, interior offers must exceed the offer at the boundaries.[7] In figure 1 we assume the wealth offers rise continuously with A from the lower boundary to a peak somewhere near the "optimal" match and

---

[7] For a proof, see Becker, Landes, and Michael (1976), an early, lengthy version of this paper.

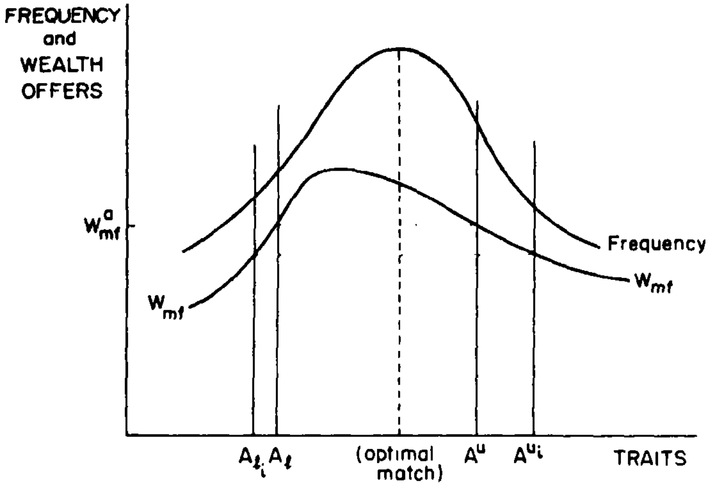


FIG. 1

then fall continuously to the upper boundary. The distribution of offers need not be symmetrical around the peak offer, so that the lower and upper boundaries will not in general be equally far from the peak.

A movement to the left along the distribution of wealth offers corresponds, although not perfectly, to a movement in either direction away from the "optimal" matching trait in figure 1. Therefore, when one accepts an offer closer to the minimum acceptable offer, he generally accepts a greater "mismatch," a greater deviation between his actual and his "optimal" matching trait. An increase in search costs alone lowers one's minimum acceptable offer and widens the boundaries of his acceptable set of traits in figure 1. Greater mismatches become acceptable because the value of additional search is reduced by the increase in search costs. Consequently, an increase in search costs can be said to increase the frequency of dissolutions because it increases the incidence of mismatches; hence, dissolutions and mismatches should be positively related empirically.[8]

In addition to "extensive" search, there is "intensive" search to improve the accuracy and reliability of expectations about a particular match. An individual spends time and other resources learning more about a potential spouse through dating and other contacts because his expectations are partly determined by information he has about himself and the potential mate. In a simple model of this search process, evidence on the match accrues at a constant rate during a courtship and after marriage. Clearly, the probability of dissolution would be smaller the smaller the variance in the distribution of realized wealth; it would also be smaller the longer the duration of a marriage because only marriages with favorable realizations survive long durations.[9]

---

[8] The equilibrium acceptable sets of men and women in the marriage market are interrelated in a very simple way. *If A_{l_i} is the lower bound of males with the trait A_{m_i}*, then A_{m_i} is the upper bound for females with A_{l_i} (for a proof, see Becker et al. 1976). Similarly, A_{m_i} is the lower bound for females with A^{u_i}. Therefore, if all the male boundaries were known, all the female boundaries would also be known, and vice versa. Moreover, *if all male boundaries increased as their trait increased, then all female boundaries would* also increase as their own trait increased.

This simple model can be generalized to permit the flow of evidence to depend on direct search outlays and on whether the search was prior or subsequent to marriage. Using the arguments developed for extensive search, we can show that an increase in intensive search costs reduces the optimal accumulation of information prior to marriage. As a result, the probability of dissolution would be greater when intensive search costs were greater because the probability of entering into a mismatch—a match involving a greater variance in outcomes and possibly a lower mean outcome—would be greater. Therefore, an increase in either the cost of intensive or extensive search would increase the probability of dissolution.

Moreover, the optimal amounts of intensive and extensive search are not independent. Presumably, a person skilled at one kind of searching also tends to be skilled at the other; also, an increase in the value of one's time increases the cost of both kinds of search. Since extensive and intensive search are positively related, smaller expected gains from marriage (due to less extensive search) and less reliable expectations (due to less intensive search) tend to go together. Consequently, the expected gain and the variance in realizations are probably negatively related, not independent, as we have been assuming.

Several determinants of the cost of search are now considered. If a matching trait is rare—such as very high or very low intelligence or an uncommon race or religion—extensive search costs would be greater because persons with average traits are more readily encountered in the marriage market.[10] That is, the frequency distribution of offers to persons looking for rare traits is less dense in the region of acceptable offers.[11] Consequently, the probability of mismatches, and thus of marital dissolutions, would be greater with rare traits.

---

[9] Jovanovic (1976) develops a model of intensive search along these lines in the context of matching employees and firms, and derives these and other implications. For a similar model of search, see Mortenson (1976).

[10] Markets are sometimes organized in ways that facilitate marital search. Examples include dances for tall persons, social activities centered around a church, residential segregation of minorities, and coeducational universities that require considerable intelligence for admission

[11] Wessels (1976) shows that the region of acceptable offers is wider and the probability of a mismatch greater when the distribution is less dense.

Women who become pregnant accidentally while searching for a mate have an incentive to marry quickly, even if they have not completed their search, because of their desire to "legitimate" their children, and because they become less valuable to other potential mates. Put differently, they are more likely to accept a mismatch because the cost to them of additional intensive and extensive search has increased. Therefore, accidental premarital conceptions should increase the probability of marital dissolution.

An important finding in practically every study of marital dissolution is that persons marrying much younger than average have significantly higher probabilities of dissolution. If the cost of search differed primarily because of differences in, say, the cost of time or even the incidence of premarital conceptions, persons with higher costs would marry relatively young and would be relatively more likely to dissolve their marriages.[12]

Age at marriage also depends on the degree of bias in expectations. Persons who are excessively pessimistic about their distribution of potential offers relative to the offers sampled (or excessively optimistic about the sampled offers relative to the distribution) tend to marry earlier because the sampled offers appear to be attractive compared to the value of additional search. Similarly, optimists about the distribution of offers (or pessimists about the sampled offers) tend to marry later because additional search appears attractive.

The additional evidence accruing after marriage would induce persons who were excessively optimistic about their mates to revise downward their expectations and would thereby increase the probability of dissolution. Since persons marrying at young ages are on average more optimistic, they would be more likely to dissolve their marriages. The probability of dissolution may not continue to decline with age at marriage, however. As persons continue to be unmarried, their expectations become more realistic, and they reduce their minimum acceptable income offers; they reduce their acceptable offers also because the number of years they could remain married would be declining. This is especially relevant for women, since after age 40 they have a limited capacity to bear children. A reduction of their minimum acceptable offers, however, raises the probability of dissolution because it reduces their gain from marriage. Therefore, the probability of dissolution could begin to rise for persons marrying at older ages.[13]

There is a common belief that dissolutions are evidence of marital failure that should be avoided if at all possible. Dissolution is a response, however, to *new* information, favorable as well as unfavorable. Moreover, the freedom to dissolve reduces the impact of unfavorable information, and thereby reduces the incentive to delay marriage or otherwise search more in order to avoid a mismatch.

---

[12] The effect of differences in search efficiency are less clear-cut. Although less efficient searchers make fewer searches, they spend less total time on search only if the elasticity of the number of searches with respect to change in efficiency is sufficiently great.

[13] The effect of a change in the opportunity cost of search is discussed in Becker et al. (1976).

4. *Dissolution and Investment in Marital-specific Capital*

Married persons invest in many assets, including houses, children, market and nonmarket skills, and information. Some of these investments, such as in household appliances, automobiles, or knowledge of consumer prices, would be almost as valuable if marriage dissolved. Other investments, however, would be much less valuable. Children are an important example of the latter type, since one parent usually has much less contact with the children after dissolution. Other examples are sexual adjustment with one's spouse, and specialized market and nonmarket skills used relatively more while married. The investments that are significantly less valuable when single can be called "marital specific" (see Becker 1974, p. 338).

The accumulation of "general" capital does not affect the expected gain from remaining married compared to dissolution, whereas the accumulation of marital-specific capital raises the expected gain because, by definition, this capital is not as valuable when single. Therefore, the accumulation of specific capital discourages dissolution.

Of course, the causation runs in both directions: the possibility of dissolution also discourages the accumulation of specific capital because such capital is less valuable after dissolution. For example, persons with high search costs, such as those with rare traits, or persons unlucky in their search would tend to invest less in children and specific skills because their marriages have a higher probability of dissolution. They may be especially cautious in the first few years of marriage when the probability of dissolution is usually higher. Indeed, a major reason why couples search intensively during the first few years after marriage is to improve their information before they invest substantially in specific capital.

Since an autonomous increase in the probability of dissolution discourages investment in specific capital, which further increases the probability of dissolution, an increase in search costs would increase this probability partly because it induces a decline in specific-capital investment (for a proof, see Becker, Landes, and Michael 1976). Moreover, expectations become self-fulfilling in the sense that a rise in the anticipated probability of dissolution may be partly realized only because the induced decline in specific capital increases the actual probability of dissolution.

Perhaps after an initial period of caution due to uncertainty about dissolution, marital-specific capital would grow with duration, at first rapidly, then more slowly, including a possible decline at long durations.

Since specific capital reduces dissolutions, the probability of dissolution would tend to decline at a decreasing rate with duration; as the stock of specific capital eventually declined—perhaps because children grew up—dissolutions might eventually even begin to increase.

5. *Dissolution and Remarriage*

Although we have assumed that persons dissolving their marriages must remain single, the great majority in the United States eventually remarry: 80 percent of divorced males and 75 percent of divorced females remarried in the 1967 SEO survey. Even countries that forbid divorce and legal remarriage cannot prevent common law or "consensual" remarriage.[14] When remarriage is possible, the wealth expected from remaining married would be compared not only to the wealth from becoming divorced, but also to that from remarrying.

Remarriage has significant effects on the timing and incidence of dissolutions. An unexpected increase in wealth—perhaps because one mate's earnings or the other's nonmarket productivity was greater than anticipated—would increase the gain from continuing the marriage compared to becoming single again because married wealth typically would be increased by more than single wealth. The probability of dissolution would be reduced, therefore, if being single were the only alternative to remaining married. If remarriage is possible, however, the probability of dissolution might well be increased because the gain from marrying someone else could increase by more than the gain from remaining married to the current mate. For example, a more educated, attractive, competent, or healthy mate would have been selected if a person anticipated that his earnings, personality, or health would turn out as well as it did. His actual mate would try to maintain their marriage by giving him a larger share of their full wealth. But beyond some point, their combined wealth from dissolution would exceed their wealth from staying together.

This positive relation between unexpected events and the incentive to dissolve marriage can be used to reconcile the actual evidence on dissolutions with some popular beliefs. For example, it is almost universally believed that higher-income persons separate and divorce more frequently than others, yet statistical studies invariably show the opposite. Since an unusually large fraction of persons who were favorably surprised have high incomes—such as persons who married as undergraduates and became successful lawyers, physicians, or executives—popular beliefs can be dominated by the positive effect of favorable surprises on dissolutions, whereas the statistical evidence is dominated by the negative effect of high anticipated ("permanent") incomes.

---

[14] See, e.g., the study by Kogut (1972) of the incidence and stability of consensual unions in Brazil.

The possibility of remarriage could greatly increase the probability of dissolution, since the realized wealth from a marriage could remain above single wealth but could be below a much higher expected wealth from remarriage. Moreover, a decrease in the expected gain from marriage compared to being single might actually reduce the probability of dissolution because the expected gain compared to remarriage could increase. For example, a reduction in the minimum acceptable marriage offer reduces the gain compared to being single but could increase the gain compared to remarriage if the distribution of offers and the minimum acceptable offer were the same in both the remarriage and the first-marriage markets.

Nevertheless, specific capital, search costs, and variables that affect the gain from marriage under certainty tend to have the same qualitative effects on the probability of dissolution when remarriage is possible as we have shown them to have when remarriage is excluded. This is obvious for the capital specific to a particular marriage—such as children—and can also be easily shown for variables like expected income and beauty that affect the gain from marriage.[15] It is less obvious for search costs because an increase in the cost reduces the value of search in the remarriage market along with the minimum acceptable offer for a first marriage.

Yet an increase in the cost of search would tend to increase the probability of dissolution even when the distribution of offers and the minimum acceptable offer were the same in both the remarriage and first-marriage markets. For one thing, if an increase in the cost of search increased the cost of intensive search (and hence the variance of outcomes from a given match) along with the cost of extensive search, the probability of dissolution would increase because a larger fraction of outcomes from a first marriage would be less than the minimum acceptable offer, which equals the value of searching for a new mate.

In addition, an increase in the cost of search increases the probability that search after marriage would reveal a preferable match. When remarriage is possible, continued marital search may be quite rational, and the frequency of extramarital relations is some evidence on the importance of such search.[16] Since an increased cost of search lowers the expected value of the offer accepted in the first marriage, it raises the probability that a random drawing from the remarriage market will produce a better match, and much of the search after marriage may be equivalent to random search in the remarriage market. To be sure, some of the offers after marriage may not be randomly chosen and may depend on the effort devoted to finding them. Since an increase in the cost of search raises the cost of this effort as well, an increase in the cost need not be positively related to the number of attractive offers. Presumably, however, the "spontaneous" or "random" search that does raise the number of attractive offers for persons with high search costs is an important part of the total search of married persons, since their marital status often severely limits the effort they can devote to search.

---

[15] A shift to the right in the distribution of offers raises the expected gain from marriage compared both to the minimum acceptable offer—which also shifts to the right but by less than the shift in the distribution—and to the "offer" from being single.

[16] Perhaps more persuasive evidence is that a significant number of persons remarry shortly after their first marriage dissolves (see the empirical evidence in table 6).

We conclude that even when the distribution of offers and hence the minimum acceptable offers are the same in both the remarriage and the first-marriage markets, couples with less marital-specific capital, higher search costs, and otherwise lower expected gains from marriage and larger variances in outcomes dissolve their first marriages more readily. Consequently, our earlier analysis that neglected remarriage is still applicable when the remarriage market is not more attractive than the first marriage market.[17]

Since divorced persons tend to have lower expected gains and higher variances in outcome from marriage than persons remaining married, the average person marrying a second time would tend to have a lower expected gain and higher variance than the average person marrying a first time.[18] Therefore, the dissolution rate on second marriages of persons divorced the first time should tend to exceed the rate on first marriages.[19] More generally, the dissolution rate and the order of a marriage should be positively related.

Specific capital can also explain why second or later marriages are more likely to dissolve than first marriages, even when duration of current marriage, age at current marriage, and other characteristics are held constant. Children (and perhaps other specific capital) from previous marriages could reduce the stability of the current marriage because they are a source of friction; that is, positive specific capital in one marriage could be "negative" specific capital in a subsequent marriage.[20] Moreover, persons who dissolved their first marriage may have anticipated dissolution and invested more in general ways that would be useful when divorced. These investments in turn reduce the stability of subsequent marriages by increasing the attractiveness of divorcing again. One implication, therefore, is that termination of a first marriage per se increases the probability of dissolving future marriages because of the destabilizing effects of specific and general capital from the first marriage.[21]

---

[17] Moreover, several arguments suggest that opportunities in the remarriage market are *less* favorable than opportunities available before the first marriage (see Becker et al 1976).

[18] This difference in the average expected gain on first and second marriages is reduced but not eliminated by the positive relation between the probability of remarriage and the expected gain from remarriage.

[19] On the other hand, there is little reason to expect persons who were widowed the first time to have a relatively high dissolution rate on their second marriage, for confirmation, see the empirical evidence in Section 11.5.

[20] In the same way, positive specific capital in one firm could *lower* the productivity of a worker moving to another firm because he has become "accustomed" to the first firm's methods and organization and has lost some of his "flexibility."

6. *Summary of Theoretical Analysis*

A list of the major implications derived from the theoretical analysis provides a useful summary for the empirical analysis in Section II.

1. An increase in the expected value of variables positively sorted in the optimal sorting of mates, such as the earnings of men and the attractiveness of women, lowers the probability of dissolution and raises the probability of remarriage if dissolved. The reason is that the expected gain from marriage will increase. On the other hand, an increase in the expected value of variables negatively sorted in the optimal sorting of mates, such as the earnings of women relative to those of men, raises the probability of dissolution and lowers the probability of remarriage given dissolution.

2. A larger deviation between actual and expected values, such as actual and expected earnings or fecundity, raises the probability of dissolution. The reason is that the gain from becoming divorced or from marrying someone else increases by more than the gain from remaining married to the same spouse.

3. An increase in education has an ambiguous effect on the probabilities of dissolution and remarriage. The reason is that education reduces the division of labor between mates (thus lowering the gain from marriage) while increasing the gain from any given division of labor.

4. An increase in age at marriage tends to reduce the probability of dissolution, especially at relatively young ages. The reason is that persons marrying relatively young have greater search costs and are less informed about themselves, their mates, and the marriage market. However, the probability of dissolution may begin to rise with age at marriage at relatively older ages.

5. An increase in marital-specific capital, exemplified by young children, reduces the probability of dissolution. The reason is that such capital would be worth less in any other marriage or when divorced.

---

[21] In the same way, separation from one job per se increases the turnover on all subsequent jobs, which can contribute to the explanation of differences in turnover rates between so-called movers and stayers. Instead of taking the differences in behavior between stayers and movers as given, our analysis probes into the underlying causes, such as differences in search costs, specific capital, properties of optimal sortings, and even luck.

Conversely, an increase in the probability of dissolution reduces the demand for marital-specific capital. Children and perhaps other specific capital may also lower the probability of remarriage and raise the dissolution rate on remarriages because they hinder the search for another mate and reduce the gain from remarriage.

6. A larger discrepancy between the traits of mates and what they would be in the optimal sorting—for example, discrepancies between intelligence, social background, religion or race—raises the probability of dissolution and lowers the probability of remarriage if divorced. The reason is that the gain from marriage is reduced. More generally, an *increase in the cost of finding a suitable mate increases the probability of* dissolution.

7. The probability of dissolution tends to decline as the duration of a marriage increases. The reason is that marital-specific capital, such as children, sexual compatibility, and knowledge of one's mate, increases with duration. The observed probability of dissolution would decline with duration in a given cohort of marriages not only for this reason, but also because couples with higher probabilities of dissolution dissolve their marriages relatively early; therefore, the average probability of those remaining married would decline even if each couple's probability were invariant with duration.

8. The speed and probability of remarriage depend directly on the expected gain from remarriage; therefore, they depend directly on variables like male earnings and inversely on variables like female earnings and the stock of capital specific to prior marriages (such as children from those marriages). Since marriages tend to last longer when the expected gain is greater, they also depend directly on the proxy variable, duration of prior marriages.

9. The probability of dissolution is higher in second than in first marriage, is still higher in third marriage, and so forth. The reason is that persons dissolving their marriages are not selected at random but are selected by characteristics that increase their probability of dissolution, such as lower gains from marriage. Moreover, even if dissolutions of first marriages were randomly selected, the dissolution rate on subsequent marriages would be greater because children and perhaps other specific capital from first marriage lower the gain from subsequent marriages.

II. Empirical Analysis

Fortunately, many of the theoretical implications listed above can be explored empirically using several bodies of data which give the incidence of divorce and remarriage by duration of marriage, number of children, education, earnings, age at marriage, number of marriages, and other variables. Indeed, since data on the stability of marriages are more extensive than are data on job or residential stability, marital behavior offers a fertile area for testing a theory about the stability of contractual relations.

We have analyzed in detail two data sets: primarily, a nationwide survey of approximately 30,000 households conducted by the U.S. Bureau of the Census in 1967 (the Survey of Economic Opportunity [SEO] data), and also a survey of approximately 1,500 persons with IQs over 135 who were first surveyed in 1921 by psychologist Lewis Terman and who were resurveyed over the subsequent 50 years (the Terman Survey). These two data sets, as well as findings from many other studies, enable us to investigate extensively the implications derived in Section I.

In Section II we first present findings on first-marriage divorce rates for men and women separately and then present some evidence on the relationship between search costs, marital-specific capital, and the probability of divorce. Next we consider the likelihood of remarriage and second-marriage divorce. We conclude with a brief discussion of secular changes in divorce rates.

I. *Stability of First Marriage*

The SEO survey contains a large amount of information on the marital histories of the men and women surveyed: the number of times married, the dates of first and current marriages, the date and type of termination of first and current marriage, and, by inference, the age at marriage.[22] For the white men and white women aged 35–55 at the time of the survey we constructed eight separate data files pertaining to the stability of each person's first marriage, in 5-year marriage duration intervals beginning with the date of marriage and running through the twentieth anniversary of the marriage.[23] The frequency of divorce and sample size of each of these files is indicated in table 1.

Men.—Let P_i be the probability of divorce in the ith 5-year interval of marriage ($i = 1, \ldots, 4$), conditional upon being married at the beginning of that interval. Then the logistic function

$$P_i = 1/1 + e^{-(\beta_i X + u_i)} \tag{1}$$

was estimated by maximum likelihood separately for $i = 1, \ldots, 4$ where

$$\beta_i X = \beta_{0i} + \beta_{1i}(AM) + \beta_{2i}(AM)^2 + \beta_{3i}(S) + \beta_{4i}(A) + \beta_{5i}(E) + \beta_{6i}(E)^2, \tag{2}$$

where AM is age at first marriage, S is years of schooling completed by 1967, A is age in 1967, and E is annual earnings in 1966. Table 2 reports

---

[22] Glick and Norton (1971, p. 308) discuss the strengths and weaknesses of the SEO data for the study of marital behavior.

[23] Persons whose marriage ended by death of a spouse were excluded; persons for whom a specific 5-year interval was interrupted by the 1967 survey were excluded from that interval. Becker et al. (1976) discuss why persons under age 35 were excluded and report empirical results for persons aged 45–54.

TABLE 1

FREQUENCY OF DIVORCE AND SAMPLE SIZE IN SEO SURVEY

	WHITE MEN		WHITE WOMEN	
	% Divorced	Sample Size	% Divorced	Sample Size
First 5 years of marriage	3.5	4,413	4.1	5,509
Second 5 years of marriage . . .	2.3	4,045	3.9	5,184
Third 5 years of marriage	2 1	3,337	3.6	4,588
Fourth 5 years of marriage . . .	1.7	2,156	2.4	3,235

TABLE 2

EFFECTS OF AGE AT MARRIAGE, SCHOOLING LEVEL, AGE, AND EARNINGS ON THE PROBABILITY OF DIVORCE, BY MARRIAGE DURATION INTERVAL (SEO White Men, Aged 35–55 in 1967)

	MARRIAGE DURATION INTERVAL (in years)			
	0–5	5–10	10–15	15–20
AM	−1.561	−.694	−.590	−.334
	(−4.59)	(−2.06)	(−1.20)	(−.33)
$(AM)^2$	.027	.013	.009	.005
	(4.38)	(1.98)	(.86)	(.22)
S	−.070	.151	−.101	.229
	(−.77)	(1.92)	(−1.21)	(2.36)
A	−.121	−.031	.082	−.034
	(−2.47)	(−.70)	(1.63)	(−.45)
E	−.334	−.181	−.177	−.310
	(−3 19)	(−1.30)	(−1.92)	(−3.01)
E^2	.004	−.000	.002	.003
	(2.74)	(−.04)	(2.32)	(1.44)
Constant	17.904	1.771	−.829	−.769
$\bar{P}$	.0351	.0227	.0210	.0172
Likelihood ratio test	46.41	12.10	20.92	13.20
N	4,413	4,045	3,337	2,156

NOTE.—The logistic function $P_i = 1/(1 + e^{-X\beta_i - u_i})$ was estimated by maximum likelihood methods, and the coefficients reported above are $\beta_i(\bar{P}_i)(1 - \bar{P}_i)$ where $\bar{P}_i$ is indicated above for each of the ith intervals. Asymptotic t-values for each β are shown in parentheses

the estimated partial derivatives of P_i with respect to each variable, evaluated at the mean $\bar{P}_i$. The same relationships were also estimated by the OLS regression

$$D_i = \beta_i X + \epsilon_i, \tag{3}$$

where D_i is 1 if the marriage dissolved in that interval and 0 if the marriage was intact at the end of the interval. The differences in the two sets of estimates were slight.

Age at marriage (AM) has a strong negative effect on the probability of divorce at early ages at marriage and a positive effect at ages of marriage beginning near 30;[24] the impact of AM is quite strong in the early years of marriage and attenuates in magnitude and significance at higher durations. A strong negative effect of age at marriage on divorce rates is one of the most widely observed correlates in the divorce literature;[25] an upturn beyond age 30 is also evidenced in census data but is not frequently recognized in the demographic literature.[26] This U-shaped effect of age at marriage on divorce is consistent with our theory (see implication 4 in Section I.6).

---

[24] The age at marriage with the minimum probability of divorce implied by the sets of coefficients are 28.9, 26.7, 32.8, and 33.4.

The effect of education on divorce is not stable in sign and generally not statistically significant. Although the simple correlations between divorce rates and education are negative, as found also in census and other data, the results in table 2[27] suggest that the effects of age at marriage and earnings explain the negative simple correlation between men's education and men's divorce rates, since these latter variables are positively correlated with education. The weak and ambiguous effect of education is not inconsistent with the theoretical analysis, for an increase in education has offsetting effects on the probability of dissolution (see implication 3 in Section I.6).

Earnings are consistently negatively related to the probability of divorce up to an earnings level of at least $40,000, and become positively related at high levels except in the second interval. Our theoretical analysis implies that a permanent increase in earnings lowers the probability of divorce,[28] and a greater deviation between actual and expected earnings increases the probability (see implications 1 and 2 in Section I.6). Since men with greater deviations in earnings are concentrated at both tails of the distribution of actual earnings, dissolutions would be especially high at the lower tail both because expected earnings are low and the deviations are large; they would then decline as actual earnings rose, but could begin to rise at the upper tail because the positive effect of large deviations could begin to outweigh the negative effect of high expected earnings. Therefore, our theoretical analysis can readily explain the initially strong negative and eventually positive relation between actual earnings and the probability of divorce.

To test this interpretation, we have reestimated the OLS regressions, replacing the variables E and E^2 by variables measuring expected earnings ($\hat{E}$) and the absolute value of unexpected earnings ($|E - \hat{E}|$).[29]

---

[25] See, e.g., Carter and Glick 1970, esp. pp. 234–35; Bumpass and Sweet 1972; or U.S. Bureau of the Census 1972*a*. Ross and Sawhill's (1975) linear coefficient implies that a 5-year delay in age at marriage, *ceteris paribus*, reduces the probability of divorce over a 4-year period by two percentage points (the average probability in the sample is 7.6 percent).

[26] The upturn at older ages is in evidence for all marriage cohorts since 1920 (see U.S. Bureau of the Census 1973, table 4).

[27] In an often cited study of 1960 census data, Cutright (1971, p. 293) also concludes that education of husband has no appreciable effect on the stability of first marriages when their earnings are held constant.

[28] It should also encourage earlier marriages, and this implication is strongly confirmed with the SEO survey (see Keeley 1974).

[29] For the white men in the SEO survey, the log of earnings was regressed on three sets of variables: (1) years of schooling, S, experience (defined as age $-$ S $-$ 6), X,

Expected earnings have negative, generally statistically significant, effects on dissolutions in each interval, and unexpected earnings have positive, but small and insignificant effects; both effects are in the predicted directions. A further analysis separates unexpected earnings into positive and negative deviations. Unexpectedly high as well as unexpectedly low earnings had positive effects on the probability of divorce in every duration interval, which adds additional support to our interpretation.

More direct evidence on the effect of unexpected events on the probability of divorce is available from other studies, and it supports our interpretation of the findings with respect to earnings. A spell of unemployment often indicates longer-run difficulties in the labor market that were not anticipated at the time of marriage. Our analysis then implies that persons experiencing extended unemployment would tend to have relatively high probabilities of divorce. Ross and Sawhill (1975, p. 56) find that men who experienced serious unemployment in the prior 3 years had a significantly higher probability of divorce over the subsequent 5 years. Similarly, an unexpected opportunity for remunerative employment in another community for only one spouse in families where both are employed would encourage divorce since the other spouse could be harmed by a move; hence migration would encourage divorce, and vice versa (see the analysis and evidence in Mincer 1977).

Fertility impairment is not easily identified prior to marriage, hence couples who experience sterility, spontaneous abortions, or stillbirths should be more likely to divorce. There is some evidence that women with relatively many fetal losses or child deaths are more likely to have married more than once.[30] Excessive fecundity is also difficult to predict; some evidence in the next section suggests that couples who have children too easily also have higher probabilities of divorce.

Individuals whose health changed significantly from what it was prior to marriage should also be more likely to divorce, since health changes are usually difficult to anticipate. According to evidence from the

---

experienced squared, X^2; (2) S, X, X^2, and marital status; and (3) S, X, X^2, and weeks worked. (Becker et al. [1976] discuss the rationale for these three equations.) Earnings expected when marrying for the first time, $\hat{E}$, is assumed to be equal to the earnings predicted from the equation above with age set equal to 45 (with no adjustment for secular growth in earnings across cohorts—see Section II.6). "Unexpected earnings" simply equals the absolute value of the difference between actual and predicted earnings at the current age. Our estimates of expected and unexpected earnings are admittedly crude, and the magnitudes of the resulting coefficients are sensitive to the use of auxiliary regression 1, 2, or 3. Becker et al. (1976) report the results of these regressions.

[30] In one recent nationwide data set, 88 percent of *currently* married women with no child deaths were married only once, compared to less than 80 percent of those with one or more deaths. Similarly, 88 percent of *ever-married* women with no fetal loss were married only once, compared to 82 percent with one or more losses. Standardizing for the age of women does not affect this basic picture. These calculations, based on the 1970 National Fertility Survey, were kindly supplied by Anne D. Williams.

NBER-Thorndike-Hagen sample, men who report their health as either better or worse than as young men are more likely to be divorced than are men who report their health has remained about the same.[31] These results cannot be explained solely by a negative effect of marital instability on health because they hold also (although less strongly) for persons whose health has improved.

Although the estimated logistic functions in table 2 have low statistical significance, many of the estimated coefficients are significant even at the .99 level of confidence. Instead of relying exclusively on significance tests, we have tried to determine in another way whether the independent variables can discriminate among persons who divorce early, later, or never. Three separate probabilities of divorce are predicted using the coefficients for each interval and the values of the independent variables for men (1) divorcing in that interval, (2) divorcing in a later interval, and (3) still maritally stable (by 1967).

These predictions are shown in table 3. As expected, they are lowest for men still first married and highest for men who divorced in that interval. The percentage differences are reasonably large, even between the two groups who were not divorced in the interval for which the regression was estimated. Therefore, the independent variables in these regressions can discriminate between the maritally more stable and less stable.

The proportion of marriages ending by dissolution declines with duration of marriage as shown above and as implied in Section I (see implications 5 and 7, Section I.6), and most explanatory variables in table 2 have their strongest effect in the first interval.[32] In separate tests, the cumulative impact on divorce over the first 20 years of marriage of age at marriage and earnings (but not age and schooling) are considerably larger when estimated from separate duration-specific equations than when estimated while constrained to be independent of duration (see Becker et al. 1976). Moreover, the decline in the probability of divorce with duration is apparently not accounted for by changes in our explanatory variables with duration (note the pattern of the intercepts).

Women.—Comparable logistic (and linear regression) models were estimated for each 5-year marriage duration interval for white women. The logistic function was

$$P_t = 1/1 + e^{-(\gamma_t Y + v_t)} \tag{4}$$

---

[31] According to data supplied by Michael Grossman from NBER-TH survey, 1.9 percent of the men whose current health equalled their health when in high school were divorced, 2.6 percent of those men whose health deteriorated were divorced, and 2.2 percent of those men whose health improved were divorced (for a discussion of these data, see Grossman 1976).

[32] Selective attrition as a cohort moves through time may bias the estimated effects from interval to interval.

TABLE 3

Conditional Predicted Probability of Divorce (SEO White Men, Aged 35–55)

Predicted Probability for Marriage Interval	Interval-specific Probability (%) Estimated for Men: Whose Marriage *Did* End in This Interval (1)	Whose Marriage Ended in a Subsequent Interval (2)	Whose Marriage Was Intact at Time of Survey (3)	Percentage Differences $\frac{(1) - (3)}{(3)}$ (4)	$\frac{(2) - (3)}{(3)}$ (5)
0–5 years	4.90	4.00	3.52	39	14
5–10 years	2.68	2.55	2.30	17	11
10–15 years	2.76	2.80	2.07	33	35
15–20 years	2.74	2.05	1.70	61	21

Note.—Estimated for three groups using marriage-duration specific maximum likelihood estimates.

with

$$
\begin{aligned}
\gamma_i Y = {} & a_{0i} + a_{1i}(AM) + a_{2i}(AM)^2 + a_{3i}(S) + a_{4i}(A) \\
& + a_{5i}(PM) + a_{6i}(C_{1i}) + a_{7i}(C_{2i}) \\
& + a_{8i}(C_{1i} + C_{2i})^2,
\end{aligned} \tag{5}
$$

where *PM* is a dummy defined as 1 if the first birth occurred less than 7 months after the date of marriage, and C_{1i} and C_{2i} are the number of children under age 6 and between 6 and 17 at the beginning of interval i.[33]

Unlike the function estimated for men, the equations for women do not include earnings (since many of these married women did not work in 1967), but they do include the number of children at the beginning of each interval and a premarital pregnancy variable. Therefore, the regressions for women do contain variables (the children variables) that explicitly measure behavior subsequent to marriage. Consequently, the coefficients in the women's regressions, unlike those in the men's regressions, measure the effect of a variable like age at marriage net of its effect on the number and ages of children.

Table 4 reports the estimated partial derivatives of equation (4) evaluated at the mean P_i for the four 5-year duration intervals for women.[34] The effects of *AM*, *S*, and *A* are generally similar to the effects estimated for men in table 2, except age at marriage retains statistical significance at higher durations, and the effect of age, although consistently negative, does not have statistical significance even in the first interval.[35]

A premarital conception has a large effect on the probability of divorce in all four intervals, although it is statistically significant only in the second interval. We argued that an accidental premarital conception increases the probability of dissolution because it raises the cost of finding a suitable mate (see Section I.3 and implication 6 in Section I.6).[36]

---

[33] The SEO survey contains information of the birthdates of up to four children (the first two and the last two) born to each woman. For women with more than four children, birth dates for the middle children were interpolated at equal intervals between the second and penultimate child. For the first marriage interval, C_{11} was defined as of the fifteenth month of the interval instead of at the beginning of the interval. Furthermore, the analysis was restricted to women whose first child was not more than 1 year old at the date of first marriage, hence C_{21} was omitted from the first interval.

[34] Again the OLS estimates are very similar and are not discussed separately.

[35] The *AM* which minimizes the probability of divorce lies between age 25 and 30 except in the second interval when the logit estimate is an anomalous age 66, compared to an OLS estimate of age 32.

[36] The probability would also be increased if the children conceived prior to marriage were fathered by someone other than the current husband. This explanation is pursued in Section II.5 in the analysis of divorces from second and later marriages. Our results on the effect of premarital pregnancy are consistent with many other studies. For example, Grabill (1976, table 9) shows with 1970 census data that the instability of marriages is

TABLE 4

EFFECTS OF AGE AT MARRIAGE, SCHOOLING LEVEL, AGE, AND FERTILITY ON THE PROBABILITY OF DIVORCE, BY MARRIAGE DURATION INTERVAL
(SEO White Women, Aged 35–55 in 1967)

	MARRIAGE DURATION INTERVAL (in years)			
	0–5	5–10	10–15	15–20
AM	−1.691	−.791	−1.634	−1.331
	(−5.65)	(−1.61)	(−3.80)	(−2.22)
$(AM)^2$	.028	.006	.030	.023
	(4.49)	(.51)	(3.06)	(1.60)
S	−.100	.027	.065	−.062
	(−1.00)	(.26)	(.62)	(−.62)
A	−.049	−.060	−.018	−.008
	(−1.07)	(−1.27)	(−.37)	(−.14)
P	1.141	2.713	.745	.664
	(1.02)	(2.87)	(.72)	(.67)
C_1	−1.130	−2.403	−2.826	−2.086
	(−1.89)	(−4.02)	(−4.99)	(−3.34)
C_2	...	...	−1.435	−1.131
			(−2.81)	(−2.76)
$(C_1 + C_2)^2$	...	.405	.366	.169
		(2.43)	(4.26)	(2.66)
Constant	13.403	5.440	11.093	10.882
$\bar{P}$	.0412	.0392	.0355	.0235
Likelihood ratio test	69.80	70.72	50.79	38.09
N	5,509	5,184	4,588	3,235

NOTE.—The logistic function $P_i = 1/(1 + e^{-X\beta_i - u_i})$ was established by maximum likelihood methods and the coefficients reported above are $\beta_i(\bar{P}_i)(1 - \bar{P}_i)$, where $\bar{P}_i$ is indicated above for each of the ith intervals. Asymptotic t-values for each β are shown in parentheses.

The number of children under age 6 has a large and statistically significant effect on the probability of divorce. A child by the fifteenth month of a marriage lowers the probability of divorce within the first 5 years of marriage by one percentage point, or by about 25 percent of the mean (holding constant the premarital pregnancy variable). The effect of young children on the probability of divorce in the second and third intervals is nonlinear: for example, in the second interval, the first child under age 6 at the fifth year of the marriage lowers the probability of divorce in the next 5 years by about 2.0 percentage points; a second child lowers it by 1.2 percentage points; third and fourth children have small marginal effects, −0.04 percent and +0.4 percent, respectively; while a fifth child appears to *raise* the probability by 1.2 percentage points! The positive effect of a relatively large number of children appears to support our theoretical prediction that a greater deviation between

---

considerably higher among women with a premaritally conceived first child: e.g., among women first married in 1965–69, the percentage married with husband present in 1970 was 85.5 percent for women without a premaritally conceived first child, 81.6 percent for women whose first birth was within 6 months of marriage, and 70.9 percent for women whose birth occurred before first marriage.

actual and expected values of a characteristic (including control over conception) raises the probability of dissolution (see implication 2 in Section I.6).

An older child (age 6–17) has a much weaker effect on the probabilities of dissolution than does a young child. Our theory implies that children reduce the probability of dissolution because they represent marital-specific capital (see implication 5 in Section I.6); the weaker effect of older children also is consistent with this implication provided younger children embody more marital-specific capital.[37]

There has been surprisingly little quantitative evidence on the effect of children on divorce rates, although the relation between children and divorce has long been recognized.[38] With detailed evidence we find large effects of children, although these effects are not linear with respect to either number or age:[39] younger children discourage divorce more than older children do and the first two children discourage divorce more than additional children do.

2. *Search Costs and the Probability of Divorce*

Evidence from several studies indicates that discrepancies in the traits of mates (relative to that implied by the "optimal" sorting) increase the probability of dissolution. For example, considerable sociological literature has, for decades, emphasized that religious differences encourage dissolution; Landis (1949) and Bell (1938) found that the probability of dissolution was about 10 percentage points higher for a person who married outside his or her religion. Differences in education and in age also appear to increase the probability of dissolution (see Levinger 1965, p. 24).

The SEO survey is not useful in studying the effects of discrepancies because no information was collected on the traits of former spouses. The Terman survey[40] does provide such information, and Michael (1976) has related the probability of divorce to the subject's age, education, and religion, and to the spouse's education and religion. Five separate dummy variables, one for each religion (including no religion), measure whether or not the Terman subject and spouse had the same religion. Results for women indicate that all five religion variables have large and statistically significant coefficients for divorces obtained early in marriage. The probability of divorce within the first 4 years of marriage is more than 20 percentage points lower when both have the same religion than when they differ. With the exception of Jewish marriages, the effects are about as large, although not as statistically significant, for the probability of divorce in the first 24 years of marriage.

---

[37] There is evidence that parents spend more time with younger children than with older children (see Leibowitz 1974; Gronau 1976; Walker and Woods 1976; and for an international comparison, Stone 1972), suggesting that the care of younger children is more marital specific. In OLS regressions on the interval between the twentieth and twenty-fifth years after marriage an additional variable, number of children over age 17, was included and had an insignificant *positive* coefficient, and the effect was largest when C_1 and C_2 were zero (i.e., when there were no younger children remaining in the family). This result perhaps suggests that parents sometimes postpone dissolution until their children are older.

[38] See Monahan (1955, pp. 446–56); Jacobsen (1959); Levinger (1965, p. 24); Carter and Glick (1970, p. 36); and Plateris (1973*b*).

[39] Ross and Sawhill (1975) use less detailed and linear measures of the effects of children, perhaps this explains why they apparently "do not find . . . the presence of children has any significant effect on [marital] stability" (p. 57).

[40] The subjects were nonrandomly selected from California elementary schools in 1921 and had an IQ exceeding 135; thus they are in the top 1 percent of the IQ distribution. (For more intensive discussions of the sample, see Terman [1925–59], or Michael [1976].)

In Section I.3 we showed that the traits of mates differ more from what they would be in the "optimal" sorting when marital search costs are larger, and that the probability of dissolution is greater when search costs are larger or when the discrepancy between the actual and "optimal" traits is greater (see implication 6, Section I.6). The findings just cited are all consistent with this implication but not with the view that persons marrying out of their religion are simply less committed to their religion, for why then would their dissolution rates be unusually high?[41]

Further evidence comes from second and later marriages. If the propensity to intermarry is partly the result of higher search costs, and if these higher costs persist into the remarriage market, divorced persons who intermarried in the first marriage should tend to intermarry in subsequent marriages and should have relatively high dissolution rates in their later marriages. The religion-intermarriage rates of Terman subjects in their first three marriages are presented in table 5. More than one-half of the Terman women and one-third of the Terman men who remarried after dissolving a first marriage with someone from a different religion, again married outside of their own religion. This is not only much higher than the fraction of religion-intermarriages in all first

---

[41] Additional evidence also suggests that persons who intermarry tend to have higher search costs and do not simply have different tastes or less luck in their search. A study of Jews in Indiana showed that the fraction intermarrying has been much greater in communities with relatively few Jews (where the cost of finding a satisfactory Jewish mate is greater) than in communities with relatively many Jews (Rosenthal 1970, table 2). There is also evidence that persons who marry relatively young are more likely to intermarry than persons who marry at average ages (Burchinal and Chancellor 1962; and Rosenthal 1963). The high dissolution and intermarriage rates of young marrieds are related: both reflect limited information. There is some evidence that premarital pregnancies are also more common when spouses differ in religion (see Christensen and Barber 1967), which suggests that persons marry out of their religion partly because of a premarital pregnancy. This evidence on intermarriage supports our interpretation of the relation between dissolution and age at marriage. Further support is provided by the relatively high rates of intermarriage of persons marrying (for the first time) over age 30, for we have argued that they also have relatively high dissolution rates because they gain less from marriage (see Burchinal and Chancellor 1962).

TABLE 5

THE FRACTION OF TERMAN SUBJECTS WHO MARRIED SOMEONE OUTSIDE OF THEIR RELIGION, BY ORDER OF MARRIAGE AND PREVIOUS BEHAVIOR (Information from 1950 Marital Histories)

CURRENT MARRIAGE	FIRST MARRIAGE	SECOND MARRIAGE		THIRD MARRIAGE	
		Married within Same Religion on 1st Marriage	Married outside Own Religion on 1st Marriage	Married within Same Religion on 2d Marriage	Married outside Own Religion on 2d Marriage
Women:					
Married someone of *same* religion	.88	.81	.44	.40	.50
Married someone of *different* religion	.12	.19	.56	.60	.50
No. of observations	486	26	9	5	4
Men:					
Married someone of *same* religion	.86	.82	.67	1.00	.67
Married someone of *different* religion	.14	.18	.33	0	.33
No. of observations	689	38	18	4	3

marriages but is also considerably higher than the fraction of persons who intermarried after dissolving a marriage with someone from the same religion. The presumption is that for this latter group search costs are not as high with respect to intermarriage as are the costs of divorced persons who intermarried the first time but are higher than those of all persons who married the first time. Divorced persons who did not intermarry the first time did have a higher rate of intermarriage the second time than did all persons marrying for the first time.[42]

We have argued that intermarriage is higher among Jews living in communities with relatively few Jews partly because they have higher costs of finding suitable Jewish mates. Put differently, being Jewish in these communities is a rare trait that raises the cost of search and, as a result, raises the discordance in traits between mates and the dissolution rate (see Section I.3). The Terman sample was selected on the basis of a rare trait, a high IQ: far less than 1 percent of the population has an IQ exceeding the average of this sample (148). The expectation from our theory is, therefore, that Terman subjects would both marry out of their IQ class (they would "intermarry" with respect to IQ) and divorce at relatively high rates, unless they were much more efficient searchers in the marriage market.

Although information on the IQs of the Terman subjects' spouses is limited, available evidence is consistent with our expectation.[43] It is not surprising that Terman subjects have had high divorce rates: 27 percent of Terman women had been divorced from their first husband by 1972. (There is a question, however, of whether that figure is relatively high compared to the appropriate control group; see Michael [1976].)

Since the wage rates of mates are negatively correlated in the "optimal" sorting (see Becker 1974), and since women typically earn less than men, a discrepancy between mates in this trait would generally take the form of an increase in the relative wage rate of the wife. Our theory implies that the dissolution rate would be higher when her relative wage rate is higher,[44] an implication supported by considerable evidence.[45]

---

[42] There is evidence that previously divorced Jews intermarry more frequently than Jews marrying for the first time (Rosenthal 1970). Moreover, the same study suggests that the relatively high intermarriage rate on remarriages of divorced persons is not a necessary consequence of remarriage, for Jewish widows do not have a high intermarriage rate on their remarriages; indeed, it is even lower than that of persons marrying for the first time!

[43] The average score on a "concept mastery" test of spouses was about one standard deviation lower than the average score for Terman subjects, even when schooling level was held constant (see Terman 1959, pp. 57–60).

[44] Even in the optimal sorting, the expected gain from marriage would be lower, and hence the probability of dissolution greater, when the wage rate of the wife relative to that of her husband was greater (see Section I.2).

[45] E.g., see Ross and Sawhill 1975, p. 56. The evidence on Terman subjects' divorces over a 20-year period is also relevant; the schooling coefficient of Terman women is significantly positive and that of their husbands significantly negative, and schooling

Interesting additional evidence comes from the effects of welfare payments on dissolutions. Welfare conditioned on the household's income is the poor woman's alimony and, like a higher wage rate for women, reduces the gain from marriage by increasing the expected income while unmarried. Consequently welfare would reduce the gain from remaining married, and, indeed, Honig (1974) finds that the fraction of both white and black households headed by females in different SMSAs is strongly related to the size of welfare payments.[46] More generally, any system of transfers in which payments depend mainly on a household's total income—be it welfare, negative income tax,[47] or aid to families with dependent children—encourages dissolutions because it compensates for the reduction in resources available to the spouses as a consequence of dissolution.

3. *Fertility and the Probability of Divorce*

In table 1 we reported a strong relation between the probability of divorce and the number and ages of children, and we presumed that the causation ran from children to marital stability. However, an exogenous increase in the probability of divorce would reduce the demand for children and for other marital-specific capital as well (see Section I.4 and implication 5 in I.6). So a negative correlation between number of children and the probability of dissolution might reflect, instead, causation from a higher probability to fewer children. A simultaneous equations model could be constructed to try to identify the causation between children and dissolution, but we decided against this strategy. Instead, we have attempted to study causation from the probability of dissolution to the demand for children by constructing a situation that largely excludes the reverse causation.

Couples with higher probabilities of dissolution tend to have less invested in marital-specific capital not only in the early years of marriage when dissolutions are more frequent, but in later years as well both because dissolution rates differ then also, and because couples would not fully compensate later for reduced investment earlier. Building on this argument, we relate the number of children of intact couples to several

---

and wage rates (not included in the regression) are positively correlated (see Michael 1976). Further evidence comes from the analysis of aggregate data. Freiden (1974) finds that the fraction of women married in different geographic units is generally lower, even with age and several other variables held constant, when their wage rates are higher relative to those of men, Santos's findings (1975) are similar.

[46] By contrast, Sawhill et al. (1975) conclude from their study with the Michigan Panel Data that "there is no evidence . . . that higher welfare benefits increase separation rates among low-income families" (p. 97), but ADFC recipiency (not level of payments) "inhibits the marriage and remarriage rates of women who head families" (p. 98).

[47] See Hannan, Tuma, and Groeneveld 1976

determinants of their probability of dissolution and to (other) determinants of their demand for children. We have restricted our analysis to couples in which the wife is over age 40 in order to measure essentially completed fertility.

From the SEO sample, the number of children ever born to white women age 40-55 with first marriage intact was regressed on a set of independent variables which includes several generally used in fertility equations and three variables intended to reflect the probability of divorce—discrepancies between spouses in race, education level, and age. Race is defined as a dummy variable equal to zero if the spouses are not of the same race and one if they are. The education and the age variables are defined as the cross-product of the education levels and of the ages of the spouses respectively; the discrepancy is greater the smaller these cross-products.[48] The race and cross-product variables are expected to have positive coefficients on fertility because smaller discrepancies in traits result in a lower probability of dissolution and thus a higher demand for children. The relevant results from this regression on 3,262 observations are as follows. With ages of husband and wife, schooling levels of husband and wife, and husband's wage and wife's age at marriage (and age at marriage squared) included in the regression, the education cross-product has a highly significant coefficient (.016, $t = 6.96$), as predicted.[49] Race also has a strong positive effect (1.00, $t = 1.90$), as predicted: racially mixed couples tend to have one less child than other couples. The age cross-product coefficient is inconsistent with our prediction, but is very small and statistically not significant (−.002, $t = 1.33$).[50]

---

[48] Let the true demand function for children be $C = \beta Z + a_0 X_h + a_1 X_w + a_2[X_h - (d_0 + d_1 X_w)]^2$, where X_h and X_w are the education (or age) levels of the husband and wife, respectively, and Z is other variables. The term $d_0 + d_1 X_w$ gives the value of X_h that is combined with X_w in the optimal sorting; hence $d_1 > 0$ if the optimal sorting has positive assortative mating. The term in brackets is a measure of the discrepancy between the optimal and the actual value of X_h; hence $a_2 < 0$ because the probability of dissolution will be higher and the demand for children smaller when the discrepancy is greater. By expansion, $C = \beta Z + (a_0 - 2a_2 d_0)X_h + (a_1 + 2a_2 d_0 d_1)X_w + a_2 X_h^2 + a_2 d_1^2 X_w^2 - 2a_2 d_1 X_h X_w + a_2 d_0^2$. Since $a_2 < 0$ and $d_1 > 0$, the coefficient of $X_h X_w$ is positive, and the coefficients of X_h^2 and X_w^2 are negative. If $d_0 \simeq 0$, the coefficients of the linear term are unaffected by any discrepancy. Since X_h^2, X_w^2, and $X_h X_w$ are highly collinear, we eliminated X_x^2 and X_w^2 from the regression; this biases the coefficient of $X_h X_w$ downward—i.e., against our hypothesis because the omitted variables are both positively correlated with $X_h X_w$

[49] Other studies have also found that the interaction between the education of mates or sometimes the interaction between husband's income and wife's education has a positive effect on fertility (see Garrison, Anderson, and Reed 1968; Kiser 1968, and Ben-Porath 1974, and also Willis 1974). The interaction between IQs also appears to affect fertility (see Garrison et al. 1968) "presumably for the same reasons [that explain the similar results of education] whatever they may be" (p. 124). We have supplied a reason: couples whose IQs or educations or other traits differ from what they would be in the optimal sorting have fewer children because they have a greater probability of dissolution. Willis (1974) argues that the interaction between education levels has a positive effect on fertility because the value of the wife's time is inversely related to the degree of interaction. This may well contribute to the explanation of the findings on education and IQ but, unlike our argument, is not relevant to the related finding that discrepancies in race and religion (discussed below) also reduce fertility.

Additional evidence is available from a regression of the number of children ever born to Terman women (who first married prior to 1940 and whose marriage was still intact in 1960) on several variables including a dummy variable defined as one if the spouses were of the same religion. That religion variable has a sizable effect: the number of children is reduced by about 0.7 if spouses have different religions, which is more than one-third of the average number of children in this sample (the coefficient's t-value = 1.83). We reported in the previous section that Terman women are much more likely to divorce when they marry someone outside of their religion, so these data also indicate that the demand for children is lower among couples with a relatively high probability of dissolution.

This section has adduced evidence of causation running from the probability of dissolution to the demand for children: a high probability reduces the demand. There is a little evidence also that the demand for other kinds of marital-specific capital is reduced as well.[51] Direct quantitative evidence, as opposed to the indirect evidence in table 4, of causation running from children to the probability of dissolution is available in the evidence below on remarriages and dissolution of second marriages.

4. *Remarriage*

Divorced persons in the United States can remarry again if they choose to, and the overwhelming majority eventually do. The SEO sample is typical; more than 75 percent of divorced men and more than 70 percent of divorced women remarried within 15 years of their divorce. The word "eventually" needs to be emphasized, however, because remarriage is far from immediate. Only 31 percent of the SEO men and 22 percent of the women remarried within 2 years of their divorce, and only 48 and 43 percent, respectively, remarried within 5 years.[52]

Our theoretical analysis implies that the probability of remarriage is greater when the expected gain from marriage is greater as a result either of lower search costs or greater gains in the "optimal" sorting (see implication 8 in Section I.6). As a test of this implication, the probability of remarriage of divorced men and women in the SEO survey was related to several measures of the expected gain. The logit results reported in table 6 pertain to the probability of remarriage by the nth year after the termination of first marriage ($n = 2, 5, 10$, and 15 in the four regressions used in table 6).[53]

---

[50] Moreover, this coefficient as well as the education coefficient, is probably biased downward because of the omission (due to collinearity) of the squares of age and education.

[51] See Becker et al. (1976) for evidence on the relation between the participation in the labor force by married women and the discrepancy between their traits and those of their husbands.

[52] Since considerable time usually elapses between separation and divorce, the time between dissolution and remarriage is much longer.

Higher earnings for men significantly increase the probability of remarriage at all four durations.[54] This is further evidence that the expected gain from marriage is increased by an increase in men's earnings (see implication 1 in Section I.6), evidence that is consistent with the findings that an increase in earnings reduces both the probability of divorce (table 2) and the age at marriage (Keeley 1974).

Persons divorced from marriages with relatively large expected gains would tend to have been married longer than other divorced persons because more time is required to accumulate a sufficient amount of adverse information to offset larger expected gains (see Sections I.3 and implication 8 in I.6). Hence the length of the first marriage can be used as a proxy for the expected gain,[55] and should be positively related to the probability of marriage. Table 6 strongly confirms this: the probability of remarriage is raised by about one percentage point for each additional year the first marriage lasted.

Education has a statistically insignificant effect on the probability of remarriage for both men and women (again, see Hannan et al. [1976] and Sawhill et al. [1975] for comparable results). These results are consistent with the weak effect of education on the probability of divorce (see tables 2 and 4), and with the implication that an increase in education has offsetting effects on the expected gain from marriage (see implication 3 in Section I.6).

Age at divorce has a generally negative effect on the probability of remarriage,[56] an effect which is stronger in magnitude and statistical

---

[53] Note that, whereas the divorce probabilities analyzed in Section II.1 are conditional or marginal probabilities for each successive 5-year interval, the probabilities of remarriage in table 6 are cumulated over the total n years from the end of the first marriage. Hence the coefficients in table 6 give the effect of each variable during the entire time span specified. Comparable OLS regressions here too yielded essentially the same coefficient estimates.

[54] We have also run these regressions replacing annual earnings by weekly earnings, and these regressions also exhibit a significant positive effect of earnings on the remarriage rate. Similar evidence is found in other studies (e.g., see Hannan et al. 1976, p. 87; and Sawhill, Peabody, Jones, and Caldwell 1976, p. 85).

[55] It is not surprising, therefore, that expected earnings of divorced men, a direct measure of the expected gain from marriage, and the duration of marriage are positively related.

[56] The quadratic terms in AD in table 6 imply that the probability of remarriage declines for women with age at divorce above age 22, 22, 25, and 20 in the four separate logit functions, respectively; for men the probability declines with age at divorce above age 33, 23, and 17 in the three logit functions, excluding the 10-year result in which the probability declines with AD at all levels.

TABLE 6

EFFECTS OF SELECTED VARIABLES ON THE PROBABILITY OF REMARRIAGE BY SEX, BY DURATION OF TIME SINCE THE END OF FIRST MARRIAGE (SEO White Men and White Women Aged 50–65 in 1967)

	DURATION OF TIME (for Men) SINCE END OF FIRST MARRIAGE (in Years)					DURATION OF TIME (for Women) SINCE END OF FIRST MARRIAGE (in Years)			
	2	5	10	15		2	5	10	15
AD	2.64	1.87	−1.05	2.01	AD	1.77	2.20	4.98	2.73
	(1.40)	(.79)	(−.32)	(.50)		(2.24)	(1.78)	(2.73)	(1.19)
$(AD)^2$	−.04	−.04	−.004	−.06	$(AD)^2$	−.04	−.05	−.10	−.07
	(1.70)	(1.13)	(.10)	(1.03)		(−3.26)	(−3.15)	(−3.83)	(−1.83)
S	.92	1.10	−.52	−1.05	S	−.36	−.21	−.08	−.78
	(1.14)	(1.14)	(−.50)	(−.92)		(−.94)	(−.38)	(−.11)	(1.03)
A	−.31	−.08	.21	1.51	A	−.74	−.32	−.09	−1.06
	(.52)	(−.11)	(.28)	(1.81)		(−2.81)	(−.83)	(−.19)	(−1.90)
E	1.21	1.70	2.54	2.95	C	−1.72	−1.56	−3.54	−.16
	(2.00)	(2.27)	(2.95)	(2.89)		(−2.18)	(−1.38)	(−2.47)	(−.10)
DIM	1.16	1.21	.90	.90	NC	5.86	30.07	44.53	46.56
	(2.45)	(2.19)	(1.54)	(1.46)		(1.29)	(3.84)	(3.72)	(3.21)
W	−8.47	−3.37	−8.16	−11.23	DIM	.78	1.09	1.49	.92
	(−1.46)	(−.50)	(−1.17)	(−1.57)		(2.76)	(2.72)	(2.92)	(1.57)
Intercept	−69.24	−51.73	27.41	−72.91	W	−11.16	−9.31	−11.16	−15.30
						(−4.40)	(−2.53)	(−2.50)	(−3.16)
					Intercept	6.40	−11.77	−47.89	59.57
P	.285	.477	.636	.755	P	.124	.294	.434	.621
Likelihood ratio test	21.17	16.80	19.80	26.08	Likelihood ratio test	91.21	111.48	102.36	74.77
N	354	310	261	216	N	991	861	684	536

NOTE.—AD is age at time of divorce; DIM is duration of first marriage in years, W is a dummy variable equal to 1 if the first marriage ended in widowhood and 0 otherwise; C is the number of children from the first marriage, NC is a dummy variable equal to 1 if there were no children from the first marriage and 0 otherwise; S, A, and E defined as in tables 2 and 4. Asymptotic t-values of logit coefficients are in parentheses significance for women. This differential effect for women is presumably partly related to the closer connection for women between age and childbearing capacity and partly to the steep decline with age in the ratio of unmarried men to women.[57]

The probability of remarriage appears to be higher for divorced persons than for widows: the widow dummy variable has a large negative effect on the probability of remarriage that is statistically significant for women. This would not be consistent with our analysis if, as seems likely, widows gain more from marriage than divorced persons; after all, persons do not usually become widowed principally because their marriage was not successful.[58]

The dummy variable distinguishing widows from divorcees implicitly assumes that they have been in the "remarriage market" equally long when the elapsed times from legal termination of their first marriages have been equal. Yet many divorced persons begin looking for another mate as soon as they separate, and some separate only *after* they have found another mate.[59] At least part of the separated time of divorcees should be included, therefore, when calculating their length of stay in the remarriage market. Since the SEO survey did not ask for the date of separation, we have reestimated the regressions underlying table 6 after simply subtracting 2 years from the date of divorce, although the separated time of most divorced persons may well exceed 2 years (see Plateris 1973*b*, pp. 15-16). The probability of remarriage in these revised regressions (not shown here) is no longer smaller for widows; indeed, the coefficient of the widow variable is usually positive, although never statistically significant.[60]

An explicit estimate of the effect of separation can be derived from the

---

[57] E.g., in the 1960 Census the number of unmarried men declined continuously with age, while the number of unmarried women declined to age 30-34 and rose thereafter. The number of unmarried women per unmarried man, 5 years older, fell until the men were age 30-34 and rose thereafter (see U.S. Bureau of the Census 1966).

[58] The selection of widows may not be completely independent of the success of their marriage because "unhappy" persons probably die more readily than "happy" ones.

[59] Evidence from the labor market indicates that many, if not most, persons find a new job before they quit or are laid off from their old one; in one data set almost all quits and about 75 percent of layoffs were reemployed with negligible unemployment (Bartel 1975, p. 39).

[60] The percentage remarrying is much higher for divorced than for widowed women and slightly higher for divorced than for widowed men when age at termination of the first marriage is not held constant. E.g., 5 years after termination of the marriage, 48 percent of divorced men and 43 percent of divorced women in the SEO survey had remarried compared to 45 percent of widowed men and only 21 percent of widowed women. The explanation is that widowed persons are generally older, and many more women are widowed than men. Since divorces occur much earlier and an equal number of men and women become divorced, the remarriage market is much better for the younger still-fecund divorced woman than for the older widowed woman. This interpretation is borne out by 1971 figures from the Census Bureau (see U.S. Bureau of the Census 1972*b*, table 1).

TABLE 7

Regressions on the Time Interval (Years) between Termination of First Marriage and Date of Remarriage for Terman Subjects Married More Than Once by 1950 and with Spouse Present, by Sex

Variable	Women	Men
Age at termination of first marriage	−.21 (−1.17)	−.03 (−.38)
No. of children, first marriage	1.02 (2.13)	−.19 (−.70)
Duration of first marriage (years)	.10 (.45)	−.43 (−2.64)
Length of separation in first marriage (in 6-month intervals)	−.45 (1.10)	−.20 (−1.47)
Widowed	−1.33 (−.69)	−2.90 (−2.29)
Constant	8.13	6.07
R^2	.13	.17
F	1.96	3.58
Sample size	42	72

Note.—t-values in parentheses.

Terman survey as it includes information about the length of separation during the first marriage. The time interval between the legal termination of the first marriage and the commencement of the second marriage for the small number of Terman subjects in their second marriage in 1950, was regressed on the length of separation, a dummy indicating how the first marriage ended (widowed = 1), and other variables used in the analysis of the SEO data. The results in table 7 indicate that widows do remarry more quickly than divorced persons – the coefficient for men is statistically significant[61]—when the length of separation and other variables are held constant. Moreover, persons do appear to use their time while separated to search for another mate: both men and women tended to remarry more quickly when they were separated longer.

Children from the first marriage significantly reduce the probability that women remarry during any given period of time since legal termination of their first marriage (table 6, right-hand side) and increase the time it takes to remarry for those who do (table 7). The evidence in table 6 suggests that the number of children may be less important than the presence of any children. Our theory does imply that children reduce the gain from remarriage because they are specific to the first marriage, and they raise the cost of searching for another mate because they raise the shadow price of the mother's time (see Section I.5 and implication 5 in I.6).[62] We say "mother's time" because the children of divorced parents usually live with their mothers. Consequently, children from their first marriage should not have much effect on the propensity to remarry for divorced men; table 7, indeed, shows that, whereas children significantly raise the duration of time to remarriage of Terman women, they have no such effect on Terman men.[63]

---

[61] The coefficients are generally less statistically significant for women partly because many fewer Terman women had remarried by 1950 (42 compared to 72 Terman men). The regression results for men are also consistent with those from the SEO survey in two other respects: the probability of remarriage is greater for men married longer the first time and is not affected by the age at which the men divorced. These Terman results for women, however, are not consistent with those from the SEO survey, perhaps because the sample size is quite small for women.

One immediate implication of this evidence is that divorced men are more likely to remarry partly, perhaps even mostly, because divorced women usually retain custody of the children. We have crudely estimated the effect of custody by comparing the probability of remarriage of SEO men in different marriage intervals with a probability predicted for all women assuming they had no children.[64] The results are quite instructive. The *actual* frequencies of remarriage 2 years after *divorce* are 31 percent for SEO men and 22 percent for SEO women. The *predicted* frequency for women, assuming they have no children from the first marriage, is 45.2 percent, considerably above the actual probability for men![65]

The causation in the observed negative relation between children and remarriage rates rather clearly runs from additional children in the first marriage to a lower probability of remarrying. This supplements the evidence in Section II.3 that there is causation running from a lower probability of dissolution to additional children. It also reinforces our contention in Section II.1 that the observed negative relationship between children and the probability of dissolution has an important component that runs from additional children to a higher probability of remaining married.

5. *Stability of Second and Higher-Order Marriages*

More than three-quarters of persons whose first marriage ends in divorce in the United States eventually remarry; many also divorce a second time. Some remarry a third time and divorce again. Using divorce and marriage records from the state of Iowa, Monahan (1958, 1959) finds

---

[62] Also consistent with this interpretation is the evidence that women with illegitimate children take longer to marry the first time than do women the same age without any children (see Berkov and Sklar 1976, table 4).

[63] The SEO survey does not provide information on the children of divorced (or widowed) men.

[64] These predictions are estimated from regressions similar to those reported in table 4 but estimated for divorced women alone.

[65] The *actual* frequency for childless women was 100 percent all 32 of the childless women remarried within 2 years of their divorce! Differences between the remarriage rates of widowed men and women are even larger than are those of divorced persons partly because widows are older and perhaps because widowed men try to remarry quickly in order to provide their children with a parent who has or will acquire child-rearing skills.

TABLE 8

REGRESSION COEFFICIENTS ON DUMMY VARIABLES INDICATING IF PREVIOUSLY MARRIED AND PREVIOUSLY WIDOWED, FROM OLS REGRESSIONS ON THE PROBABILITY OF DIVORCE, BY SPECIFIC INTERVALS, BY SEX (SEO White Men and Women, Age 15–65)

	MARRIAGE INTERVAL (in Years)			
	Women		Men	
EXPLANATORY VARIABLE	0–5	5–10	0–5	5–10
Dummy = 1 if second or third marriage*	.138 (15.94)	.012 (1.36)	.036 (4.13)	.013 (1.68)
Dummy = 1 if widowed in first marriage	.002 (.13)	−.018 (1.19)	−.009 (.47)	−.009 (.51)
R^2 (entire regression)	.037	.010	.011	.001
F (entire regression)	56.82	12.23	12.08	.80
N	11,960	9,627	8,688	6,948

NOTE.—t-values in parentheses. Other variables included in the regressions are age, education, age at current marriage, for men their 1966 earnings, and for women the number of children from their current marriage measured at the beginning of each interval.

* Second or third marriage for men, second marriage for women.

that the probability of divorce increases sharply with the order of marriage for persons previously divorced[66] but not for persons previously widowed.

We ran OLS regressions on the probability of divorce with the SEO data including higher order as well as first marriages. The independent variables in these regressions duplicate those shown in table 2 for men and table 4 for women, but the samples differ in that we have pooled experiences on second marriages with those on first marriages for the women and have pooled experiences on second and third marriages with those on first marriages for the men.[67] In these pooled regressions, two dummy variables were added. The first indicates a previous marriage (defined as one if the observation pertains to a second or third marriage), and the other indicates a previous widowing (defined as one if the first marriage ended in widowhood). Table 8 gives the coefficients on these two dummy variables only, taken from the full multiple regression equation.

The main findings of Monahan and others continue to hold after standardization. For women, second marriages are much more unstable than first marriages, especially during the first 5 years of marriage:[68]

---

[66] In 1953–55, there were 17 divorces per 100 first marriages, 35 per 100 marriages with both spouses previously divorced once, and a whopping 79 per 100 marriages with both previously divorced at least twice (Monahan 1958, table 5).

[67] For persons in their third marriage, the SEO survey did not ask how or when their second marriage terminated. We were able to include men in their third marriage by assuming that all were divorced (rather than widowed) from their second marriage, and that their second marriage terminated during its first 5 years if 10 years elapsed from termination of the first marriage to commencement of the third, during the second 5 years if 15 years had elapsed, and so on. Women in their third marriage were excluded because a significant fraction of them were presumably widowed from their second marriage.

[68] As constructed, the first dummy variable's coefficient shows the effect on the probability of dissolution of being previously divorced compared to being in the first marriage,

the probability of divorce is about 14 percentage points higher on the second than on the first marriage However, in the first 5-year interval the probability of divorce for widows is greater than for women in their first marriage.[69]

The results for men in table 8 are similar: second and third marriages of divorced men are more unstable than first marriages of men, again especially during the initial years. Widowers are less likely to divorce after remarriage than are men previously divorced. Again, however, only in the first interval the probability of divorce is greater for widowers than for men in their first marriage.

Our theory implies that previously divorced persons gain, on average, less than others from subsequent marriages (see Section I.5 and implication 9 in I.6). Since the selection of widows is more independent of their gains from first marriage (see the evidence in the previous section), we expect marriages containing previously widowed persons to be more stable than those containing previously divorced persons.[70]

We also ran OLS regressions with the SEO data on the propensity to divorce from second marriages alone, using independent variables similar to those used for first marriages (see tables 2 and 4). Since few persons had divorced from a second marriage (for example, only 13 men divorced within the first 5 years of their second marriage), the statistical significance

---

and the sum of the two dummy variables' coefficients shows the effect on the probability of dissolution of being previously widowed compared to being in the first marriage The standardizations for age at current marriage, age and especially duration of current marriage were decisive in these findings for men and women The explanation is mainly that persons in first marriages were generally married longer and thus had more opportunity to divorce *sometime* during their marriage.

[69] The behavior of the Terman women is also consistent with these results By 1972, when they were about 60 years old, 27 percent had been divorced from their first husbands. More than 55 percent of the women who divorced the first time and remarried had divorced again—about twice the divorce rate from first marriages compared to 38 percent of the (just eight) women who were widowed the first time and had remarried Only 12 women were married a third time. Forty percent (four) of those (10) who had been divorced from both previous marriages were divorced again, whereas neither of the two previously widowed women had divorced

[70] The theory also predicts that the average duration to divorce of those terminating their marriage should decline with marriage order because the expected gain from marriage tends to be smaller for persons previously divorced twice than for those divorced only once, and still smaller for those divorced three times, and so on. This can explain Monahan's evidence (1959, table 11) of a significant decline with marriage order in the average duration to divorce for persons previously divorced but not previously widowed It can also explain the positive relation between the duration of the first marriage and the probability of remarriage (see tables 6 and 7), and the evidence in table 8 that the probability of divorce on second and third marriages is especially high during the first few years of marriage. However, when the SEO data are standardized for age and age at current marriage (Monahan's data were not standardized for these variables), the average duration is no longer related to marital order. Consequently, when appropriately standardized, the SEO data do not support our prediction that duration to divorce will decline in higher-order marriages.

of most coefficients is quite low. Yet, the signs of the coefficients are generally consistent with those for first marriages. For example, an increase in the earnings of men seems to reduce their propensity to divorce on second as well as first marriages, again except perhaps when earnings are quite high.

One interesting new result for women is that children from a prior marriage appear to increase the probability of dissolution from the current marriage,[71] whereas children from the current marriage appear to decrease this probability in second marriages, just as they do in first marriages (cf. table 4). Our explanation of both effects is that children are marital-specific capital: children from the current marriage increase and children from prior marriages decrease the gain from the current marriage (see implication 5 in Section I.6).

The positive effect of children from prior marriages is further evidence of causation from children to marital stability, since an exogenous increase in the probability that a second marriage will dissolve would hardly *raise* the demand for children in the first marriage. There is, however, also further evidence of causation from marital instability to fewer children: there are fewer children in higher order than in first marriages, even after age at current marriage and duration of current marriage are held constant. Since the probability of dissolution increases with marital order, the number of children would decline with order if the probability affects the demand for children.

6. *The Secular Trend in Divorce*

The number of divorces has grown remarkably during the last 125 years in all Western countries that permit divorce. For example, only two (!) divorces per year were granted in England between 1800 and 1850, whereas in the past year or so there have been approximately 1 million divorces a year in the United States. Based on 1973 data it is estimated that about 40 percent of new marriages in the United States will end in divorce (see Preston 1975). The reported divorce rate in the United States (the number of divorces in the year per 1,000 married women age 15 and over) rose from 4.1 in 1900 to 8.0 by 1920 to 8.8 by 1940 to 9.2 in 1960, with sizable fluctuations, especially around World War II. Since the mid-1960s, the rate of increase has accelerated; by 1970 the divorce rate was 14.9 and by 1974, 19.3. We believe that the theoretical and empirical analyses in the previous sections can contribute significantly to an understanding of these trends and fluctuations, but here we only sketch out the main considerations.[72]

---

[71] The coefficients of the number of children from the first marriage are +.44, +.83, and +.52 in the first three 5-year-duration intervals, respectively. This positive effect of children from prior marriages could even be underestimated because some of the effect may be picked up by the premarital conception variable, which has a significant, positive coefficient in the first two marriage duration intervals. A premarital conception is defined as any birth subsequent to the legal termination of the first marriage and prior to the seventh month of the second marriage, so it might capture the effect of children from "prior marriages" if some of these births were conceived during the first marriage or conceived by someone other than the second husband during the time interval between marriages. When the children from prior marriages and the children conceived from premarital conceptions are combined into a single variable of children born prior to the second marriage, the coefficient is sizable and positive in the second and third intervals (t-values = 3 and 1, respectively) but negative in the first interval (its t-value, however, is less than 1). The positive effect of the premarital conception variable in the first marriage regressions (see table 4) may also be partly measuring the destabilizing effect of children from "prior" unions. Note also the evidence that women with illegitimate children not only take longer to marry (see n. 65 above) but are also more likely to divorce than are other women (see Berkov and Sklar 1976).

The number of children per family has been declining since the beginning of the nineteenth century in the United States, and the decline accelerated during the last 20 years. Our analysis implies both that a decline in the number of children increases the probability of divorce and that an increase in this probability reduces the demand for children; the survey evidence analyzed in previous sections has confirmed that both directions of causation are important. Presumably, both directions of causation also are at work in the secular decline in fertility and secular rise in divorce. Note, however, that the recent accelerated decline in fertility began in the 1950s, at least 5 years before the accelerated increase in divorce.

An increase in the wages of women would reduce the gain from marriage, even when the wages of men increased at the same rate, because the sexual division of labor between market and nonmarket activities would decrease, and more married women would enter the labor force. Therefore, the secular growth in wages, which contributed significantly to the growth in the labor force participation of women, especially married women, probably also contributed significantly to the growth in divorce rates. Again causation probably flows both ways: divorced women (and women who anticipate divorce) have higher wages because they spend more time in the labor force.

Legal access to divorce became much easier during the last 100 years in the United States, Great Britain, and most other Western countries. Although we believe this trend toward easier divorce has been mainly a response to the increased demand for divorce, it may also have been responsible for a small part of the growth in divorce. Whatever the causation, the ease of obtaining a divorce and the fraction of women married are positively correlated across states, even after age and many other variables are held constant (see Freiden 1974 and Santos 1975).

One of the most provocative trends in recent years is the rise in illegitimacy despite the improvements in contraceptive technology and easing of legal restrictions on abortions. According to our evidence on the effect of "prior" children, illegitimacy should raise the probability of divorce. However, the rise in illegitimacy may be responding to the same forces that raised the divorce rate, namely, both may be responses to a decline in the gains from marriage.

---

[72] Michael has begun a systematic analysis of the trend during the last 2 decades.

A growth in the divorce rate itself encourages additional divorces because the remarriage market is better when there are more divorced persons available. There is evidence in table 6 that the remarriage market did improve as the divorce rate grew over time, for the probability of remarriage during the first few years after a divorce also grew over time. Moreover, the sharp acceleration in divorce rates that began in the 1960s may have been partly caused by the prior growth in divorce rates.

Tables 2 and 4 provide some evidence on the trend in probability of divorce. The estimated trend is measured by the coefficient on age, but its interpretation across marriage intervals is subject to several qualifications (see Becker et al. 1976). While nearly all the estimated trends are positive, the only significant one suggests an increase in the probability of divorce of about 1 percent per decade over the time span covered by the SEO data (1920s to early 1960s). These estimates of the trend in divorce rates are *net* of standardizations for trends in age at marriage, years of schooling, earnings of men, and the number and ages of children born to women. These standardized estimates may be biased because standardizing with differences across households in earnings, education, or age at marriage does not correctly provide for the effect of secular changes in these variables.[73] Moreover, these estimates have not been corrected for the effect of changes over time in the earnings of women, divorce laws, the size of the remarriage market, and other variables that contributed to the observed trend in divorce rates.

---

[73] Even though an increase in male earnings or age at marriage significantly reduces the probability of divorce when comparing different households at the same moment in time, the relation between the secular increase in divorce rates and the secular changes in these variables is less clear. An increase in the earnings of one man relative to the earnings of other men in the marriage market increases his gain from marriage partly because he is able to attract a woman with more desirable attributes. On the other hand, when the earnings of all men increase, with little change in the distribution of the attributes of women, all men cannot be sorted with more desirable women. Consequently, an increase in the earnings of any man would have a smaller effect on his gain from marriage and thus on his probability of dissolution when the earnings of other men also increase. A similar argument can be made for general increases in education levels, and related arguments can be made for a decline in the average age at marriage. Therefore, the large secular growth in male earnings may not have greatly reduced, and the secular decline in age at marriage may not have greatly increased, the propensity to divorce.

The appropriate way to standardize for differences in earnings also partly depends on the life cycle in earnings. If, for example, a 35- and a 45-year-old person had the same earnings in 1966, the younger person would generally have the higher earnings profile because earnings tend to increase between ages 35 and 45.

7. *Summary of Empirical Analysis*

Many of the more important theoretical implications are listed in Section I.6. The empirical analysis using the 1967 SEO and 1920-60 Terman data, as well as the findings in many other studies, strongly support these implications and are also of interest in their own right. The main empirical findings are as follows.

1. An increase in the expected earnings of men reduces the probability of dissolution on first marriages, raises the speed and probability of remarriage if the first is dissolved, and reduces the probability of dissolution on second or higher-order marriages. An increase in the expected earnings of women, on the other hand, has the opposite effects: it appears to raise the probability of dissolution and to reduce the propensity to remarry.

2. Unanticipated events, whether favorable or unfavorable, tend to destabilize marriage. For example, either unexpectedly high or low levels of earnings of men or difficulties in conceiving children appear to raise the probability of divorce.

3. An increase in the number of children, especially younger children, from a first marriage reduces the probability of dissolution of that marriage, and the speed and probability of remarriage for mothers with custody. Indeed, if divorced women did not usually receive custody, their propensity to remarry would not be less than that of divorced men Although children from second and higher-order marriages also lower the probability of dissolution in these marriages, children from first marriages apparently raise the instability of subsequent marriages

4. An increase in the probability of dissolution, as measured in our empirical work by the propensity to marry outside of one's religion, race, or education or IQ class, reduces the demand for children and for other marital-specific capital, such as skills highly specialized to the nonmarket sector. Findings 3 and 4 together indicate that the observed negative relation between the probability of divorce and children involves causation running in both directions.

5. A person who marries outside of his or her religion is much more likely to dissolve the marriage, to marry out of his religion if he does remarry, and then to divorce again. Moreover, even if a divorced (but not a widowed) person married in his religion the first time, he is rather likely to marry outside his religion the second time. The propensity to marry outside of one's religion, and then to dissolve the marriage, is partly related to the number of potential mates of the same religion that is available.

6. Persons who marry relatively young are far more likely to dissolve their marriages than are those who marry at "normal" ages. This has been well known, but less well known is our finding that persons who

marry for the first time relatively late—for example, in their early thirties—have the highest probabilities of dissolution.

7. The propensity to remarry is positively related to male earnings, the absence of young children, the length of time separated before legal termination of the first marriage, and the duration of the prior marriage (a variable that serves as a proxy for unmeasured determinants of the expected gain from marriage). Widowed men or women are more likely to remarry than are divorced women or men, after allowance is made for age at legal termination and some other variables.

8. The probability of dissolution is much higher on second marriages, and still higher on third marriages, for persons previously divorced but not for persons previously widowed.

Most of our empirical evidence involved different households at a moment in time. Yet a limited examination of evidence on trends in divorce rates suggests that our theory can also contribute significantly to understanding the secular growth in divorce, including the acceleration which began in the 1960s. The most important variables appear to be the decline over time in number of children, the growth in labor force participation and earnings power of women, the growth in the breadth of the remarriage market as more persons become divorced, and perhaps also the growth in legal access to divorce, illegitimacy, and public transfer payments.

III. Concluding Remarks

We recognize that the statistical significance of many empirical findings from the SEO survey is weak, especially in view of the usual amount of experimentation involved in empirical economic research. Yet we believe that the empirical evidence in Section II as a whole tells a consistent story that offers significant support for the theory developed in Section I. The evidence is adduced from many data sets, analyzed by many different researchers, and encompasses individuals in quite different income, education, religion, and family background categories. The data sets include a large cross-sectional survey with oversampling among low-income families, a large, more representative sample that follows persons for several years, a small sample of "geniuses" followed for 50 years, and other data including a small sample of Jewish and Catholic marriages, a sample of women in California with illegitimate children, and samples from the divorce registry in Iowa. The determining variables in the various analyses conducted or referenced here include earnings, education, and health of men and women, age at marriage, duration married, "intermarriage" by religion, race, age, and education, illegitimate children, and legitimate children from first and later marriages, and divorces from second and third as well as first marriages.

It is easy to develop special theories to explain specific findings on marital dissolution, and many have been suggested. It is a challenge to find a single theory that adequately explains the rich set of findings reported in Section II. The theory advanced in Section I, based on utility maximization under uncertainty by participants in marriage and remarriage markets, passes this test reasonably well. We suspect that the theory will do even better when it is more fully developed, with the significant remaining gaps filled in. These gaps include a derivation of the equilibrium sorting of mates when there is uncertainty about traits and a more satisfactory treatment of the interactions between the possibility of remarriage, optimal extensive and intensive search, and first marriage divorce.

The approach to marital dissolution developed here should also prove useful in analyzing the dissolution of (implicit as well as explicit) contracts of indefinite duration between employees and employers, business partners, friends, etc. The case for a common theoretical approach to all social behavior would be greatly strengthened if the same theory is applicable to employee turnover and the termination of friendships, as well as to marital dissolutions.

References

Bartel, Ann P. "Job Mobility and Earnings Growth." NBER Working Paper no. 117, November 1975.

Becker, Gary S. "A Theory of Marriage." In *Economics of the Family*, edited by T. W. Schultz. Chicago: Univ. Chicago Press, 1974.

Becker, Gary S.; Landes, Elisabeth M.; and Michael, Robert T. "Economics of Marital Instability." NBER Working Paper no. 153, October 1976.

Bell, H. M. "Youth Tell Their Story." Washington: American Council on Educ., 1938.

Ben-Porath, Yoram. "Economic Analysis of Fertility in Israel." In *Economics of the Family*, edited by T. W. Schultz. Chicago: Univ. Chicago Press, 1974.

Berkov, Beth, and Sklar, June. "Does Illegitimacy Make a Difference? A Study of Life Chances of Illegitimate Children in California." *Population and Development Rev.* 2 (June 1976): 201–17.

Burchinal, Lee G., and Chancellor, Loren E. "Ages at Marriage, Occupations of Grooms and Interreligious Marriage Rates." *Social Forces* 40 (May 1962): 348–54.

Bumpass, Larry, and Sweet, James. "Differentials in Marital Instability: 1970." *American Sociological Rev.* 37 (December 1972): 754–66.

Carter, Hugh, and Glick, Paul C. *Marriage and Divorce: A Social and Economic Study*. Cambridge, Mass.: Harvard Univ. Press, 1970.

Christensen, Harold T., and Barber, Kenneth E. "Interfaith versus Intrafaith Marriage in Indiana." *J. Marriage and Family* 29 (August 1967): 461–69.

Cutright, Phillips. "Income and Family Events: Marital Stability." *J. Marriage and Family* 33 (May 1971): 291–306.

Evans, Bergen. *Dictionary of Quotations*. New York: Delacorte, 1968.

Feldstein, Martin. "Temporary Layoffs in the Theory of Unemployment." *J.P.E.* 84, no. 5 (October 1976): 937–58.

Freiden, Alan. "The United States' Marriage Market." In *Economics of the Family*, edited by T. W. Schultz. Chicago: Univ. Chicago Press, 1974.

Garrison, Robert J.; Anderson, V. Elving; and Reed, Sheldon C. "Assortative Marriage." *Eugenics Q.* 15 (June 1968): 113–27.

Glick, Paul C., and Norton, Arthur J. "Frequency, Duration, and Probability of Marriage and Divorce." *J. Marriage and Family* 33 (May 1971): 307–17.

Grabill, Wilson H. *Premarital Fertility.* Bureau of the Census, Current Population Reports, Special Studies Series P-23, no. 63. Washington: Government Printing Office, August 1976.

Gronau, Reuben. "The Allocation of Time of Israeli Women." *J.P.E.* 84, no. 4, pt. 2 (August 1976): S201–S220.

Grossman, Michael. "The Correlation between Health and Schooling." In *Household Production and Consumption*, edited by N. E. Terleckyj. New York: Nat. Bur. Econ. Res., 1976.

Hannan, Michael; Tuma, Nancy; and Groeneveld, Lyle P. "The Impact of Income Maintenance on the Making and Breaking of Marital Unions: Interim Report." Research Memorandum 28, Stanford Res. Inst., June 1976.

Hashimoto, M. "Wage Reduction, Unemployment and Specific Human Capital." *Econ. Inquiry* 13 (December 1975): 485–509.

Honig, Marjorie. "AFDC Income, Recipient Rates, and Family Dissolution." *J. Human Resources* 9 (Summer 1974): 303–22.

Jacobson, Paul H. *American Marriage and Divorce.* New York: Rinehart, 1959.

Jovanovic, Boyan. "An Equilibrium Model of Labor Turnover and Unemployment." Mimeographed. Univ. Chicago, May 1976.

Keeley, Michael. "A Model of Marital Formation: The Determinants of the Optimal Age at First Marriage and Differences in Age at Marriage." Ph.D. dissertation, Univ. Chicago, 1974.

Kiser, Clyde V. "Assortative Mating by Educational Attainment in Relation to Fertility." *Eugenics Q.* 15 (June 1968): 98–112.

Kogut, E. L. "The Economic Analysis of Demographic Phenomena: A Case Study for Brazil." Ph.D. dissertation, Univ. Chicago, 1972.

Kuratani, M. "Specific Training, Employment Stability and Earnings Distribution in Japan." Ph.D. dissertation, Columbia Univ., 1973.

Landis, Judson T. "Marriages of Mixed and Non-mixed Religious Faith." *American Sociological Rev.* 14 (June 1949): 401–7.

Leibowitz, Arleen. "Home Investments in Children." In *Economics of the Family*, edited by T. W. Schultz. Chicago: Univ. Chicago Press, 1974.

Levinger, George. "Marital Cohesiveness and Dissolution: An Integrative Review." *J. Marriage and Family* 27 (February 1965): 19–28.

Martin, Donald. "Economics of Employment Termination Rights." *J. Law and Econ.*, vol. 20 (April 1977).

Michael, Robert T. "Factors Affecting Divorce: A Study of the Terman Sample." National Bureau of Economic Research Working Paper no. 147, August 1976.

Mincer, Jacob. *Family Migration Decisions.* Mimeographed. Columbia Univ., August 1977.

Monahan, Thomas P. "Is Childlessness Related to Family Stability?" *American Sociological Rev.* 20 (August 1955): 446–56.

———. "The Changing Nature and Instability of Remarriages." *Eugenics Q.* 5 (June 1958): 73–78.

———. "The Duration of Marriage to Divorce: Second Marriages and Migratory Types." *Marriage and Family Living* 21 (May 1959): 134–38.

Mortenson, Dale T. "Turnover, Tenure, and Specific Capital." Mimeographed. Northwestern Univ., February 1976.

Plateris, Alexander A. *Divorces: Analysis of Changes.* Vital and Health Statistics, Series 21, no. 22, National Center for Health Statistics. Washington: Government Printing Office, 1973. (*a*)

———. *100 Years of Marriage and Divorce Statistics: 1867–1967.* Vital and Health Statistics, Series 21, no. 24, National Center for Health Statistics. Washington: Government Printing Office, 1973. (*b*)

Preston, Samuel H. "Estimating the Proportion of American Marriages that End in Divorce." *Sociological Methods and Res.* 3 (May 1975): 435–60.

Rosenthal, Erich. "Studies of Jewish Intermarriage in the United States." *American Jewish Yearbook* 64 (1963): 3–53.

———. "Divorce and Religious Intermarriage: The Effect of Previous Marital Status upon Subsequent Marital Behavior." *J. Marriage and Family* 32 (August 1970): 435–40.

Ross, Heather L., and Sawhill, Isabel V. *Time of Transition: The Growth of Families Headed by Women.* Washington: Urban Inst., 1975.

Santos, Fredricka P. "The Economics of Marital Status." In *Sex, Discrimination, and the Division of Labor,* edited by Cynthia B. Lloyd. New York: Columbia Univ. Press, 1975.

Sawhill, Isabel; Peabody, Gerald E.; Jones, Carol A.; and Caldwell, Steven B. *Income Transfers in Family Structure.* Washington: Urban Inst., 1975.

Stone, Philip J. "Child Care in Twelve Countries." In *The Use of Time,* edited by A. Szalai. The Hague: Mouton, 1972.

Terman, Lewis M. *Genetic Studies of Genius.* Vols. 1–5. Stanford: Stanford Univ. Press, 1925, 1926, 1930, 1947, and 1959, respectively.

U.S. Bureau of the Census. *U.S. Census of the Population: 1960. Marital Status.* Subject Report PC(2)-4E. Washington: Government Printing Office, 1966.

———. *Social and Economic Variations in Marriage, Divorce and Remarriage: 1967.* Current Population Reports, Series P-20, no. 223. Washington: Government Printing Office, 1971.

———. *U.S. Census of the Population: 1970. Marital Status.* Subject Report PC(2)-4C. Wahsington: Government Printing Office, 1972. (*a*)

———. *Marriage, Divorce and Remarriage by Year of Birth: June, 1971.* Current Population Reports, Series P-20, no. 239. Washington: Government Printing Office, 1972. (*b*)

———. *U.S. Census of the Population: 1970. Age at First Marriage.* Subject Report PC(2)-4D. Washington: Government Printing Office, 1973.

Walker, Kathryn E., and Woods, Margaret E. *Time Use: A Measure of Household Production of Family Goods and Services.* Washington: American Home Econ. Assoc., 1976.

Wessels, Walter J. "The Theory of Search in Heterogeneous Markets: The Case of Marital Search." Ph.D. dissertation, Univ. Chicago, 1976.

Willis, Robert J. "Economic Theory of Fertility Behavior." In *Economics of the Family,* edited by T. W. Schultz. Chicago: Univ. Chicago Press, 1974.

Location of Firms within a Square Central Business District

Donald J. O'Hara

University of Rochester

In a square central business district (CBD), with rectilinear (east-west and north-south) travel, production is assumed to require face-to-face transactions between firms. Each firm's locational decision therefore reflects travel costs between that location and *all* other CBD locations. The assumption that firms are uniformly distributed permits an analytical solution: rent contours are circular and office rent declines as the square of the *radial* distance from the center. Thus, in contrast to travel-only-to-the-center models, the rent function is concave from below; this property is shown to hold for arbitrary firm distributions. Endogenous firm distributions, investigated numerically, generate nearly circular rent contours.

> Our intense appreciation of the fact that time is money has made the cost of land in our large cities extraordinarily great; for it is, of course, the wish to save time that makes us crowd our places of business so closely together.
> [VanRensselaer 1967, p. 95]

Production within an urban area, especially within the central business district (CBD), is often said to be characterized by "agglomerative economies." These are external to the firm economies of aggregate scale that result when each firm, having access to the great diversity of specialized products and services available within the CBD, can afford to be highly specialized itself (Hoover 1968; Segal 1976). To realize these

The author is indebted to Walter Oi for detailed comments on an earlier version, to Nicholas Rau for a suggestion that led to the explicit solution in Section III, to members of the University of Rochester Labor Workshop, and to two anonymous referees. Financial support was provided under ONR contract N00014-68-A-0091.

[*Journal of Political Economy*, 1977, vol 85, no. 6]
© 1977 by The University of Chicago. All rights reserved.

economies, there must exist a complex network of transactions among firms, each of which has many suppliers and customers, with transactions between any pair constituting a small and variable fraction of their total transactions (Vernon 1960, pp. 68–86, 100–106).

This paper develops a model of the firm's locational decision when CBD production has the above characteristics, and it investigates the model's implications for office- and land-rent patterns within the CBD. The model assumes that transactions can only occur face to face and therefore require travel by highly skilled people whose time is valuable; transportation of intermediate and final products may also be required. Hence, in choosing a location each firm will consider the travel costs between that location and *all* other CBD locations. In contrast, a worker choosing a residential location traditionally considers travel costs only between that location and a single destination, the edge (or center) of the CBD. The multiplicity of destinations considered by each firm is what makes the land-allocation problem within the CBD qualitatively different from that outside the CBD.

The complex network of transactions is represented below by the assumption that transactions between a given firm and each other firm are equally likely.[1] Of course, a firm does not actually choose suppliers and customers randomly; instead, it balances travel costs against the advantages of obtaining a quality or specialty more precisely suited to its needs. However, if a firm requires many inputs, then even an optimal choice of location, given the locations of potential suppliers, will leave the firm with substantial travel costs to suppliers located in all directions.[2] More important, a firm with highly variable input requirements will not find it economical to relocate every time that changes in these requirements lead to a change in suppliers. Such a firm's transactions may therefore be approximately characterized by travel to a representative sample of all CBD locations, although the assumption of strictly equiprobable transactions no doubt overstates the amount of travel between each firm and the more distant locations within the CBD.

The diversity of products and services available within the CBD is an instance of Smith's principle that the (economical extent of the) division of labor is limited by the extent of the market. It is apparently a property of technology that given resources often produce a larger *quantity* of output when devoted to producing a single product than when divided among several.[3] Only when the market is sufficiently large, however, will the *value* of the larger quantity of specialized output exceed the value of the smaller total quantity of more varied outputs. To represent these economies of aggregate scale, the model below avoids explicit treatment of product diversity and focuses instead on the value of output. It is assumed that all firms produce the same output (dollars) but that the amount (or value) of the output obtained from given inputs is larger, the larger the number of firms in the CBD.[4]

---

[1] A somewhat similar assumption is used in Solow and Vickrey's analysis of a long, narrow city (1971).

[2] Where a single supplier or customer dominates a firm's transactions, the pair will obviously coordinate their location decisions and are best considered as a single firm for the purposes of the present analysis.

[3] See the formulation in Alchian (1959).

The remainder of the paper is organized as follows. Section I presents a very simple model incorporating the above assumption and compares its implications with those of models assuming travel only to the center of the CBD. In this model, identical firms are uniformly distributed over the entire CBD. Sections II, III, and IV investigate the consequences of allowing nonuniform firm distributions, first examining arbitrary distributions then endogenously determined distributions. Section V returns to the basic model and relaxes the assumption that all firms are alike. Section VI contains some concluding remarks.

I. The Basic Model

A simple model of locational decisions within the CBD can be constructed on the following assumptions:

1. There are F identical firms, each occupying a unit of office space and each producing the same amount of output per period by initiating N transactions with other firms.[5] Each transaction requires a separate trip,[6] and all other firms are equally likely to be the other participant in the transaction. The amount (or value) of output obtained from these transactions is a function of the number of firms in the CBD: $q = q(F)$; $q_F \geq 0$.

2. The CBD is square, with sides of length S. All travel is north-south and east-west. This departure from the traditionally circular cities with radial and perhaps circumferential travel is both realistic and, it happens, mathematically convenient.[7]

---

[4] Once all firms have been assumed to produce the same output, a literal interpretation of the model would suggest that each firm should conduct all its transactions with the firm next door. Hence, an explicit assumption dictating that transactions take place with a variety of firms in different locations is required.

[5] Each firm also participates in an average of N transactions per period that are initiated by other firms. A firm is assumed to incur travel costs only on the transactions it initiates.

[6] Note that if each transaction requires a separate trip, the assumed firm-to-firm travel is less costly than requiring all transactions to take place in a central market. (If firms are in opposite quadrants, firm-to-firm travel and meeting in the center are of equal cost; for other pairs of firms, firm-to-firm travel is cheaper.) Central markets thus economize on finding (high-priced) buyers and (low-priced) sellers only at the cost of additional travel.

[7] As will become apparent, the square, though tractable, is not an equilibrium shape for the CBD. On the endogenous determination of the CBD's size and shape, see n. 9 below.

3. The F firms are uniformly distributed over the CBD. A uniform density is clearly neither optimal nor realistic. In effect, this assumption rules out the substitution of capital for land that is a central feature of urban land-use patterns. The consequences of relaxing this restriction are investigated in subsequent sections.

4. The travel occasioned by transactions produces no congestion.[8] The round-trip cost of such travel is C per unit distance. The absence of congestion costs is a convenience, of course, but it can also be defended as a reasonable first approximation. Congestion in the typical CBD appears to be much less, if not exactly zero, during the business hours when transactions-travel takes place than during rush hours when workers are commuting to and from the CBD. This observation suggests that only a fraction of workers can be engaged in transactions-travel at one time, with others remaining in the office. Moreover, it must be primarily the more highly paid workers who engage in transactions. If low-wage messengers could be used instead, it would pay to decentralize business activity throughout the surrounding residential area; this would save the bulk of commuting costs for office workers. The transactions of the legal, financial, advertising, and other services that locate in downtown areas seem to conform reasonably well to the above picture.

On these assumptions, a firm located at (x, y) will incur an average travel cost per transaction of

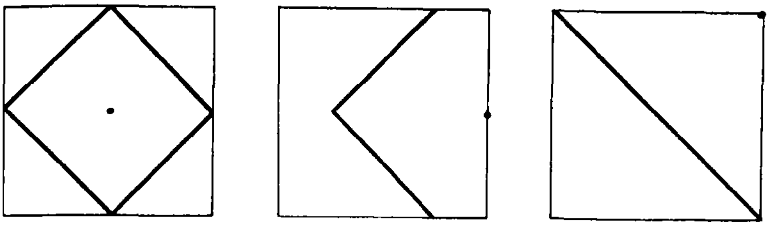
$$\begin{aligned} T(x, y) &= \frac{C}{S^2} \int_{-S/2}^{S/2} \int_{-S/2}^{S/2} |x - u| + |y - v| \, du \, dv \\ &= \frac{C}{S^2}\left(Sx^2 + \frac{S^3}{4} + Sy^2 + \frac{S^3}{4}\right) = \frac{CS}{2} + \frac{C}{S}(x^2 + y^2). \qquad (1) \end{aligned}$$

Thus, although transactions take place through rectilinear travel, the isocost contours are circles. Travel cost increases as the square of a firm's *radial* distance from the center of the CBD.

Equation (1) gives C times the average distance traveled per transaction by a firm located at (x, y). For a few locations, this is easily illustrated. For example, at the center $(0, 0)$ the average distance is $S/2$; at the middle of an edge $(S/2, 0)$ it is $3S/4$; at a corner $(S/2, S/2)$ it is S. Figure 1 illustrates these three cases. In each diagram the heavy line represents those locations at the average distance, and every destination closer than the average has a symmetrically located counterpart the same distance further away than the average.

---

[8] Since there are no congestion costs, there are no trade-offs between land use for roads or for firms, and it is implicitly assumed that travel uses no land. Such trade-offs are investigated by Miyao (1976); eq. (2) below becomes a special case of his equation 12 if rent in the corners of the CBD is constrained to be zero.



(a) Firm at (0,0) (b) Firm at (S/2,0) (c) Firm at (S/2,S/2)

FIG. 1.—Destinations at the average travel distance for firms in three locations

Since the value of output is independent of location, the bid-rent profile will be that value minus the cost of transactions-travel. Each firm initiates N transactions per period, so the office-space rent function is

$$R(x,y) = q(F) - NT(x,y) = q(F) - \frac{NCS}{2} - \frac{NC}{S}(x^2 + y^2). \quad (2)$$

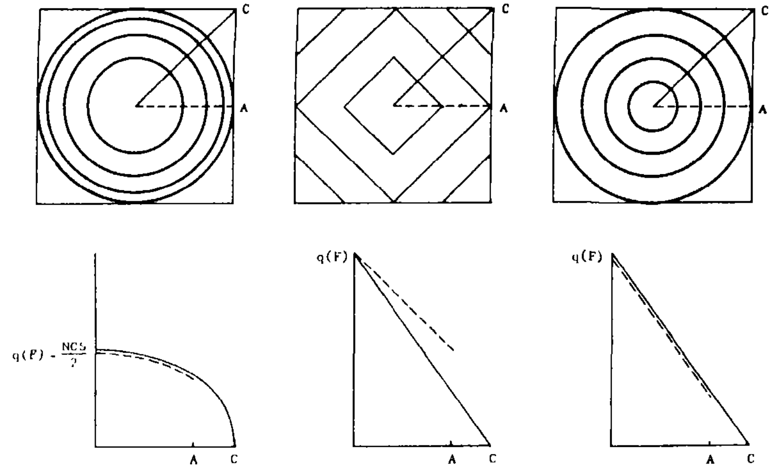
Thus, the isorent contours are also circles, with rent declining as the square of the radial distance from the center of the CBD, as shown in figure 2*a*.

The land-rent function will be $W(x,y) = G \cdot R(x,y) - K(G)$, where $G = F/S^2$ is the density of firms paying office rent and $K(G)$ is the cost of providing that density of office space. Thus, with a uniform density of firms, the land-rent function is just a linear transformation of the office-rent function. The slope and curvature of the latter determines that of the former.

For purposes of comparison, consider a model in which rent is determined by the cost of travel to the center of the CBD, retaining the rectilinear street pattern and the assumption of no substitution between capital and land. The latter model implies a rent function $R^c(x,y) = q(F) - NC(|x| + |y|)$. Isorent contours are diamonds, with rent declining in strict proportion to the east-west plus north-south distance from the center of the CBD, as shown in figure 2*b*.

Empirical investigations of rent profiles often find that radial distance performs reasonably well as an explanatory variable, despite the presence of rectilinear streets. The present model provides one explanation for this result, in that travel costs depend on radial distance to the center even though travel to the center is no more important than travel to any other point in the CBD. The implied rent function is, however, quadratic rather than linear in radial distance.

Alternatively, the explanatory power of radial distance could be the result of a world in which travel to the center is in fact the important thing and in which street widths and traffic lights have therefore been



(a) Rectilinear travel to all locations

(b) Rectilinear travel to the center

(c) Radial travel to the center

FIG. 2 - Isorent contours and rent profiles for alternative models

adjusted to facilitate such travel, with the result that radial distance is a better measure of travel time than rectilinear distance. Still assuming no capital-land substitution, such a world would generate the rent function $R^r(x, y) = q(F) - NC(x^2 + y^2)^{1/2}$ with circular rent contours but a linear rent profile, as shown in figure 2*c*.

Thus, the quadratic rent profile seems to be a distinctive feature of the present model. One could easily introduce a production technology that required both travel to all locations and a significant amount of travel to the center, perhaps to "export" that portion of output not sold to other firms in the CBD. The rent function would then have both linear and quadratic terms. Of course, the precise rent function obtained depends on the assumptions made. In particular, travel-to-the-center models produce a linear rent function only when no capital-land substitution is permitted; with such substitution, the rent function becomes convex from below (e.g., a negative exponential [see Mills 1969; Muth 1969]). It is thus of interest to discover whether the rent function in the travel-to-all-locations model retains its concave-from-below character when capital-land substitution is permitted.

II. Travel Cost and Office Rent with an Arbitrary Distribution of Firms

In actual cities, capital-land substitution obviously produces a nonuniform distribution of firms. Section III investigates the equilibrium distribution of firms and the resulting office-rent and land-rent functions when capital-land substitution is permitted. However, because buildings are extremely durable, this equilibrium will be achieved only in the correspondingly long run. Indeed, with continuing changes in construction and transportation technology, long-run equilibrium might never even be closely approached.

This section therefore considers the short-run problem of determining the office-rent function, given a fixed distribution of office space and hence of firms. It is shown below that the office-rent function is concave from below for an arbitrary distribution of firms. The results of this section thus apply not only to departures from long-run equilibrium but also to CBDs with irregular boundaries and with pockets of land devoted to nonoffice uses such as parks, churches, and parking lots.

For notational convenience, the CBD is still embedded in a square with sides of length S; the function describing the distribution of firms, $G(x, y)$, takes on the value zero wherever there are no offices. The total number of firms is then $F = \iint G(x, y)\, dx\, dy$, and the average travel cost for a firm at (x, y) is

$$T(x, y; G) = \frac{C}{F}\int_{-S/2}^{S/2}\int_{-S/2}^{S/2} (|x - u| + |y - v|)G(u, v)\, du\, dv. \quad (3)$$

The resulting office-space bid-rent function is

$$R(x, y; G) = q(F) - N \cdot T(x, y; G). \quad (4)$$

Clearly, the slope and curvature of the rent function depend on those of the travel-cost function; for expositional simplicity, only the latter is manipulated below.

The slope of the travel-cost function is, after canceling some terms:

$$\frac{\partial T(x, y; G)}{\partial x} = \frac{C}{F}\int_{-S/2}^{S/2}\left[\int_{S/2}^{x} G(u, v)\, du - \int_{x}^{S/2} G(u, v)\, du\right] dv. \quad (5)$$

Breaking the right-hand side of (5) into two parts and interchanging the order of integration produces two expressions, the first equaling the number of firms to the west of x ($-S/2$ to x) and the second equaling the number of firms to the east. Now consider the effect on average travel cost of moving a unit distance eastward from x. The distance to each firm west of x increases by one unit, and the distance to each firm east of x decreases by one unit. The net increase in travel cost is therefore proportional to the difference between the number of firms to the west and to the east.

In particular, when half the firms are on either side of x, $\partial T/\partial x = 0$, which suggests a natural choice of the east-west origin; the *median* longitude. If the north-south origin is chosen analogously, travel cost will be a minimum, and rent a maximum, at $(x, y) = (0, 0)$.

Note also that the right-hand side of (5) does not contain y; the change in travel cost when moving east-west from a given longitude does not depend on one's north-south position. Thus, the travel-cost function is separable; it can always be written as the sum of two functions, one depending only on x and the other only on y.

Consider next the curvature of the travel-cost function:

$$\frac{\partial^2 T}{\partial x^2} = 2\frac{C}{F}\int_{-S/2}^{S/2} G(x, v)\, dv. \tag{6}$$

The integral on the right is the density of firms on the north-south line passing through x. As one moves from west to east at x, these firms are both subtracted from the group of firms that is getting closer and added to the group that is getting farther away. They thus contribute twice to increasing the rate at which travel cost increases.

Because the right-hand side of (6) is always positive, assuming there are any firms at all along the longitude through x, travel cost increases at an increasing rate as one moves away from the origin. Since the office-space rent function has a slope and curvature that are opposite to those of the travel-cost function, it is concave from below for essentially any distribution of firms.

III. Capital-Land Substitution and Endogenous Firm Distributions

One firm's use of land near the center imposes higher travel costs on other firms which are precluded from using that land, and these costs are reflected in its rent. This provides an incentive to economize on these costs by substituting other inputs for land, especially by building taller buildings in the center than on the periphery of the CBD.

The extent to which it pays to build up at a particular location depends on the saving in travel cost at that location relative to other locations. In the present model, however, the travel cost in a given location depends on the density of firms at all other locations. Thus, the equilibrium distribution of firms and building heights cannot be determined point by point but instead must be determined for the whole CBD simultaneously. It is easy enough to write down the integral equation that the equilibrium distribution must satisfy, but discovering its properties is more complicated. Limited results can be achieved using the results of the previous section, and an analytical solution can be obtained for a special case, but otherwise resort to numerical solutions obtained with a computer appears necessary.

Assume that increased density of office space can be provided at increasing marginal cost, with the same cost function applicable to all locations. In particular, let $K(G) = aG^b$, $b > 1$, which is the cost function corresponding to a Cobb-Douglas production function in which capital has the exponent $1/b$ and land the exponent $1 - 1/b$. Office space will be provided at each location up to the level at which the bid rent for office space equals the marginal cost of office space, or

$$K'[G^*(x, y)] = R(x, y; G^*); \tag{7a}$$

$$ab[G^*(x, y)]^{b-1}$$

$$= q(F^*) - \frac{NC}{F^*} \int_{-S/2}^{S/2} \int_{-S/2}^{S/2} (|x - u| + |y - v|) G^*(u, v) \, du \, dv, \tag{7b}$$

where G^* is the equilibrium distribution of firms and R^{G^*} is the equilibrium office-rent function. Equilibrium land rent, $W^*(x, y)$, is simply total office rent less the cost of providing that office space:

$$W^*(x, y) = G^*(x, y) \cdot R(x, y; G^*) - K[G^*(x, y)]. \tag{8}$$

Elementary considerations (or the results of the previous section) imply that both G^* and W^* decline as one moves away from the origin. On the more interesting question of their curvature, differentiate (7) twice and use $\partial R/\partial x = (b - 1)baG^{*b-2} \, \partial G^*/\partial x$ to obtain, omitting asterisks for clarity:

$$(b - 2)(b - 1)baG^{b-3} \frac{\partial G^2}{\partial x} + (b - 1)baG^{b-2} \frac{\partial^2 G}{\partial x^2} = \frac{\partial^2 R}{\partial x^2}; \tag{9a}$$

$$\frac{b - 2}{G} \frac{\partial R}{\partial x} \frac{\partial G}{\partial x} + \left(\frac{\partial R}{\partial x} \bigg/ \frac{\partial G}{\partial x}\right) \frac{\partial^2 G}{\partial x^2} = \frac{\partial^2 R}{\partial x^2}; \tag{9b}$$

$$\frac{\partial^2 G}{\partial x^2} = -\frac{b - 2}{G} \left(\frac{\partial G}{\partial x}\right)^2 + \frac{\partial G}{\partial R} \frac{\partial^2 R}{\partial x^2}. \tag{9c}$$

From the previous section, $\partial^2 R/\partial x^2$ is always negative, so (9c) is unambiguously negative when $b \geq 2$. Unfortunately, this is at or above the upper limit of plausible values for b since it implies an exponent of 0.5 or less for capital in the production function for office space. For $b < 2$, G^* will be concave from below near the origin, where $\partial G/\partial x$ is small and both $\partial G/\partial R$ and $\partial^2 R/\partial x^2$ are large (both are increasing functions of G, which has its maximum at the origin). Away from the origin, the distribution of firms may become convex from below.

Much the same qualitative result applies to land rent, W^*. Substituting (7c) into (8), collecting terms, and differentiating gives:

$$W = (b - 1)aG^b; \tag{10a}$$

$$\frac{\partial W}{\partial x} = (b - 1)baG^{b-1} \frac{\partial G}{\partial x} = (b - 1)R \frac{\partial G}{\partial x}; \tag{10b}$$

$$\frac{\partial^2 W}{\partial x^2} = (b - 1) \left[\frac{\partial R}{\partial x} \frac{\partial G}{\partial x} + R \frac{\partial^2 G}{\partial x^2}\right]. \tag{10c}$$

The sign is ambiguous for any $b > 1$, but concavity is more likely near the origin where $\partial R/\partial x$ and $\partial G/\partial x$ are small, R is large, and $\partial^2 G/\partial x^2$ is negative. The land-rent function will, however, be convex whenever the distribution of firms is convex and, in fact, will become convex while G^* is still concave.

An explicit solution for G^* can be obtained when $b = 2$. The first term in (9a) becomes zero and G^{b-2} disappears from the second term; the latter effect implies that the separability of the travel-cost function carries over to G^*, and (9a) can be written, letting $G^*(x, y) = H(x) + H(y)$:

$$2a\frac{\partial^2 H}{\partial x^2} = -\frac{2NC}{F^*}\int_{-S/2}^{S/2}[H(x) + H(v)]\,dv$$

$$= -\frac{2NCS}{F^*}H(x) - \frac{2NC}{F^*}\int_{-S/2}^{S/2}H(v)\,dv; \qquad (11a)$$

$$H'' + B^2 H = Q, \qquad \text{where } B^2 = \frac{NCS}{aF^*} \text{ and } Q = -\frac{NC}{aF^*}\int_{-S/2}^{S/2}H(v)\,dv. \qquad (11b)$$

A solution is given by $H(x) = c_1 \cos Bx + c_2 \sin Bx + Q/B^2$. Since $H'(0) = 0$, c_2 must be zero, and the equilibrium distribution of firms is

$$G^*(x, y) = c_1(\cos Bx + \cos By) + 2Q/B^2. \qquad (12)$$

Integration yields $Q = -NC/2aS$; F^* and c_1 are found by solving $F^* = \iint G^*$ and $2a[H(S/2) + H(S/2)] = q(F^*) - NCS$. The latter equation is (7b) evaluated at the corner of the CBD; the average travel distance for a firm in the corner is always S for any symmetric distribution of firms (that is, fig. 1*c* always applies).

V. Numerical Solutions with Endogenous Firm Distributions

To investigate the properties of the model when $b \neq 2$, it was necessary to obtain numerical solutions using an iterative procedure and a computer. The procedure is simple and would be nearly as easy to apply to irregular CBDs as to the squares considered here. One chooses values for N, C, a, and b; a function $q(F)$; and some initial distribution of firms, G^0. One then calculates the travel-cost and office-rent functions implied by this distribution, using equations (3) and (4), and "builds up" the CBD to the heights that are warranted by this rent function according to the marginal cost condition in equation (7a). This produces a new distribution of firms, G^1, for which a new office rent can be calculated. The process is repeated until the difference between G^i and G^{i-1} is arbitrarily small.

This iterative procedure may be of economic as well as computational interest. A landowner planning to erect a multistory building might determine its height by balancing construction costs against the office rent currently being obtained at that location, which is determined by the current distribution of firms (or building heights). This decision rule corresponds precisely to one step in the iterative procedure. Unfortunately, although a single landowner who follows this rule will have his office-rent expectations fulfilled, if all owners follow the rule the distribution of firms changes, the rent function changes, and all expectations are disappointed. The degree of "irrationality" embodied in this mechanism depends on the rate at which the stock of office space changes and on the responsiveness of the rent function to changes in the distribution of firms, a factor investigated below.

A limitation of numerical investigations is that they produce only a collection of examples. Compared with an analytical solution, even the most carefully constructed set of examples produces a less complete understanding of which assumption leads, by what process, to a given characteristic of the solution. With that caveat in mind, numerical solutions for a variety of parameter values are examined below for the light that they shed on the following questions: (1) How far do the isorent contours depart from circularity when the firm distribution is endogenous? (2) Under what conditions will the land-rent function remain concave from below throughout the CBD? (3) How rapidly does the iterative procedure approach equilibrium, or, roughly, how irrational is the expectations mechanism embodied in the procedure?

A natural "origin" for parameter values is provided by the fact that the average travel distance for a firm in the corner is always S, independent of the distribution of firms (provided that the distribution is symmetrical). If firms are to be viable in the corner, $q(F)$ must be at least NCS; by setting $q(F) \equiv NCS$, $NC \cdot 2S$, $NC \cdot 3S$, etc., the response of the model to different levels of $q(F)$ can be investigated without becoming involved in the consideration of marginal agglomerative economies. Further, multiplying both $q(F)$ and NC by any factor just changes the scale, as does any change in a; hence both NC and a can be set equal to unity. That leaves the parameter b, the exponent in the office-space cost function; ($1/b$ is the capital elasticity in a Cobb-Douglas production function).

Table 1 reports a variety of characteristics of the model's solution when $q(F)$ is identically equal to S, $2S$, $3S$, and $4S$, and when b equals 2.00, 1.50, 1.25, and 1.10. For both parameters, the middle pair of values is most plausible and/or interesting. When $q(F)$ equals S, both office rent and land rent are zero in the corner; this is hardly characteristic of actual CBDs, which are located on the most valuable land in the urban area. At the other extreme, when $q(F)$ equals $4S$, the equilibrium firm

TABLE 1

NUMERICAL SOLUTIONS

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
$q(F) = S$:									
$b = 2.00$....	.87	.07	.35	.57	.77	.51	.51	2	.98
$b = 1.50$....	.77	.14	.45	.43	.57	.26	.51	3	.88
$b = 1.25$....	.62	.27	.63	.30	.40	.07	.51	4	.53
$b = 1.10$....	.37	.53	.92	.18	.23	.00	.51	5	.04
$q(F) = 2S$:									
$b = 2.00$....	.97	.01	.27	.99	...	.83	.83	2	1.00
$b = 1.50$....	.94	.03	.30	.74	...	.69	.83	2	1.00
$b = 1.25$....	.87	.07	.36	.55	.72	.46	.82	3	.96
$b = 1.10$....	.67	.22	.57	.33	.43	.11	.80	4	.62
$q(F) = 3S$:									
$b = 2.00$....	.98	.00	.26	...	...	.90	.90	2	1.00
$b = 1.50$.	.96	.01	.28	.99	...	.81	.90	2	1.00
$b = 1.25$....	.93	.03	.31	.72	.96	.65	.90	2	.99
$b = 1.10$....	.81	.12	.42	.43	.57	.30	.89	3	.89
$q(F) = 4S$:									
$b = 2.00$....	.99	.00	.26	...	...	.93	.93	2	1.00
$b = 1.50$....	.97	.00	.27	.	...	.86	.93	2	1.00
$b = 1.25$....	.95	.02	.29	.84	...	.74	.93	2	1.00
$b = 1.10$....	.86	.08	.36	.52	.72	.45	.92	3	.96

NOTE.—Col 1 = average travel distance for a firm in the center as a fraction of that travel distance when the distribution of firms is uniform (= $S/2$) Col. 2 = deviation of rent contour at ($S/2$, 0) from circularity; equals zero with uniform distribution of firms, equals one in the limit when all firms are at the center. Col 3 = fraction of firms in the central quarter of the CBD. Col 4 = fraction of the (diagonal) distance to the corner at which the land-rent function ceases to be concave. Col. 5 = fraction of the distance to the edge ($S/2$, 0) at which the land-rent function ceases to be concave. Col 6 = ratio of firm density (building height) at the edge ($S/2$, 0) to firm density in the center. Col. 7 = ratio of office rent at the edge to office rent in the center Col 8 = number of iterations (building-up cycles), when the initial distribution of firms is uniform, needed to achieve the "equilibrium" land-rent function; i e , a land-rent function that differs from the preceding one by less than 1 percent Col 9 = fraction of equilibrium land rent at the center achieved by the first round of building upward.

distributions for all values of b tend to approach the uniform distribution, making the analytical solution of Section I a reasonable approximation. Similarly, $b = 2.00$ implies an implausibly low capital elasticity of 0.5 and, in addition, tends to produce near uniform equilibrium firm distributions. On the other hand, $b = 1.10$ produces an implausibly large response of building height to changes in office rent; a doubling of office rent produces a 1,000-fold increase in building height. For $b = 1.5$ and $b = 1.25$, the corresponding response is fourfold and 16-fold, respectively.

The first three columns of table 1 bear on the extent to which the distribution of firms departs from the uniform and the extent to which the rent contours depart from circles. Column 1 reports the average distance traveled by a firm at the center of the CBD as a fraction of $S/2$. With a uniform distribution, this is unity (see fig. 1*a*); it declines as firms become more concentrated in the center. (The decline is always greatest at the center and is zero in the corner.) Column 2 is a test of the circularity of the rent contours. It reports (minus) the difference between the average travel distances of a firm at the end of an axis ($S/2$, 0) and a firm the same radial distance from the center along a diagonal ($\sqrt{2}S/4$, $\sqrt{2}S/4$), as a fraction of the maximum possible difference between the two. This maximum occurs in the limit when all firms are at the center (see fig. 2*b*); then the travel distance for a firm at ($S/2$, 0) is $S/2$ and that for a firm at ($\sqrt{2}S/4$, $\sqrt{2}S/4$) is $\sqrt{2}S/2$. If the average travel distance at the two points is the same, the rent contours are circular. Whatever the degree of noncircularity at the edge, it will decrease as one moves closer to the center (unless all firms are literally at the center), since the closer a location is to the center the more there are significant numbers of firms in all directions from it. Column 3 is a deliberately somewhat crude index of the degree to which firms are concentrated in the center; it reports the fraction of all firms that are located in the central quarter of the CBD. This is meant to be useful in formulating rules of thumb for predicting the behavior of the variables in columns 1 and 2. For example, in a CBD containing 40 percent of its firms in the central 25 percent of its area (the average density in this 25 percent is twice as great as that in the remaining area), one would expect only minor departures from circular rent contours. Even with 50 percent of the firms in the central quarter of the CBD, the departure from circularity might well be submerged by noise in the data.[9]

Columns 4–7 bear on the concavity of the land-rent function. Columns 4 and 5 report the fraction of the distance to the corner and to the edge, respectively, at which the land-rent function ceases to be concave. Clearly, concavity throughout is more likely the higher are $q(F)$ and b, but it is by no means guaranteed, especially for the whole length along the diagonal to the corner—even for plausible parameter values. Columns 6 and 7 again present somewhat crude indicators of CBD characteristics for use in formulating rules of thumb for predicting the curvature of the land-rent function. Column 6 gives the ratio of firm density at the edge ($S/2$, 0) to firm density at the center; column 7 gives the ratio of office rent at the same two points (see eq. [10c]). For example, if density declines by no more than one-third along an axis and rent declines by no more than 15 percent, W will probably be concave along the whole length of the axis.

The final two columns bear on the "rationality" of the expectations mechanism embodied in the iterative procedure. Column 8 reports the number of iterations required, with G^0 taken to be the uniform distribution, to obtain successive land-rent functions that differed by no more than 1 percent (i.e., the iteration whose number appears in column 8 added at most 1 percent to the land-rent function). Column 9 reports the ratio of land rent in the center after the first round of building up to the final land rent there. For plausible parameter values, the first round of building captures easily 90 percent of the final land rent and very often nearly 100 percent. The irrationality of the expectations mechanism is therefore apparently not very great.[10]

---

[9] In a variant of the model, the CBD boundary was determined endogenously by setting G equal to zero at every point where land rent fell below a specified level (presumably the maximum rent that workers would be willing to pay to live adjacent to the CBD). Although the absence of firms in any part of the square affects rent contours in the remainder somewhat, the departure from circularity is about the same as that for the corresponding cases in table 1. For plausible parameter values, therefore, the boundary of the CBD is approximately circular, and it should be possible to integrate the present model with existing analyses of residential land use, which typically assume an annular residential area encircling the CBD.

A consistent feature of table 1 is that, for a given $q(F)$, differences in b result in quite different equilibrium firm distributions, G^*; note the variation, within each panel, of columns 3 and 6. These differences in G, however, *do not*, in turn, produce great differences in the equilibrium office-rent function, R^*; note the stability of column 7, within each panel. This suggests that the simultaneous relationship between R and G is nearly recursive; G is determined by R (and by the nature of the cost function for office space), but R is not much affected by G. Thus, empirical investigations of R should not founder on irregularities in the observed distribution of firms.

V. Elaborations on the Basic Model

Relying on the (near) recursive structure of the relationship between R and G, this section returns to the uniform firm distribution of Section I and investigates the properties of R when there is variation among firms in some respect. Each source of variation is considered separately and only in its simplest form.

A. *Variation in Transactions-Travel Intensity*

Since the parameters N and C enter the office-rent function (2) only as a product, variation in either will have the same implications. Suppose that there are two types of firms, denoted with subscripts 1 and 2, and suppose that $N_1C_1 > N_2C_2$. Firms of type 1 will be said to have a higher transactions-travel intensity than firms of type 2. Obviously, this can result from being either in a line of business that requires more transactions per period or in one that involves a higher cost per unit of distance traveled. The primary travel cost in the CBD is presumably not freight

---

[10] When there are marginal agglomerative economies, both the theoretical standard of rationality and the evaluation of the iterative procedure become more complex. In this case, $q(F)$ changes after each round of building as F changes, and the iterative procedure is chasing a moving target. When $q(F) = gF^h$, with $h = 0.05$ or 0.10, convergence usually requires several iterations more than in table 1, although the changes in the last few iterations are small.

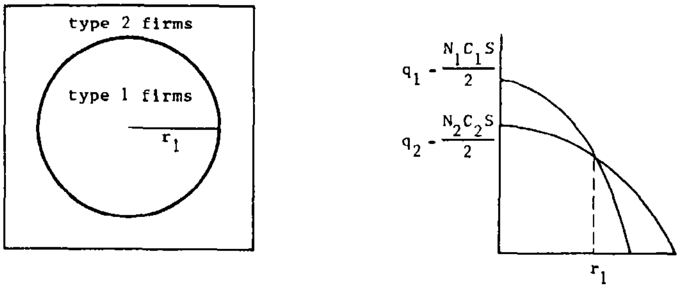


FIG. 3.—Division of the CBD between firms of high transactions-travel intensity (type 1) and firms of low transactions-travel intensity (type 2).

but the time taken by the people doing the traveling; firms that employ highly paid professionals are therefore likely candidates for the role of type 1 firms.

It is still assumed that each firm transacts with, and hence derives specialization benefits from, the presence of all other firms—not just firms of its own type. If type 1 firms are to survive, their higher travel costs must be offset by a higher value of output, $q_1(F) > q_2(F)$. Further, the bid-rent function of type 1 firms will have a steeper slope at every point than that of type 2 firms; hence, if type 1 firms are to be viable anywhere, it must be at the center. By the same logic, type 2 firms must be viable in the corners, if anywhere. Both types will be viable if

$$(N_1C_1 - N_2C_2)S > q_1(F) - q_2(F) > (N_1C_1 - N_2C_2)\frac{S}{2}. \quad (13)$$

The expression on the left is the travel-cost differential in the corner, and that on the right is the differential in the center. As illustrated in figure 3, type 1 firms occupy an inner circle of radius r_1, and type 2 firms occupy the periphery.

In a more general world, one containing viable firms along the whole continuum of transactions-travel intensity, the observed rent profile would be the envelope of bid-rent profiles. The envelope would retain the concave-from-below property suggested in figure 3, and there would be a continuous decline in the transactions-travel intensity of firms as distance from the center increased.

B. Does It Pay to Transact with the Whole CBD?

Firms have, up to this point, been required to transact with all other firms. Under certain conditions, however, it may pay a firm to forego some of the benefits of specialization in order to avoid some of the costs of transacting with the entire CBD.

Again, it is assumed for simplicity that there are only two possibilities: firms can transact either with the entire CBD or with one-quarter of the CBD. The bid-rent function of the whole-CBD firms is as before. Quarter-CBD firms, however, can choose the area within which they transact so as to minimize their travel costs; in the analysis below, that choice is subject to the constraint that the transactions area be square and oriented congruently with the CBD.

A quarter-CBD firm located anywhere within the central square with sides of length $S/2$ (A in figure 4*a*) can choose a transaction area centered on its own location. The bid rent of quarter-CBD firms will be constant within the area marked A:

$$R_{1/4A}(x, y) = q(F/4) - \frac{NC}{2}\frac{S}{2}; \qquad |x|, |y| \le S/4.$$

In the areas B and D in figure 4*a*, a quarter-CBD firm can choose its transaction area so that it is off center in only one direction; thus,

$$R_{1/4B}(x, y) = q(F/4) - \frac{NC}{2}\frac{S}{2} - \frac{NC}{S/2}\left(|x| - \frac{S}{4}\right)^2; \qquad \begin{matrix} |x| > S/4, \\ |y| < S/4, \end{matrix}$$

and similarly for $R_{1/4D}$, reversing the roles of x and y. Finally, in the areas marked C, a firm's transaction area will be its quadrant, and its bid-rent function will be the quarter-CBD analog of the whole-CBD function,

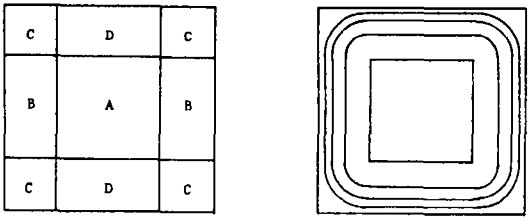
$$R_{1/4C}(x, y) = q(F/4) - \frac{NC}{2}\frac{S}{2} - \frac{NC}{S/2}\left[\left(|x| - \frac{S}{4}\right)^2 + \left(|y| - \frac{S}{4}\right)^2\right];$$

$$|x|, |y| \ge \frac{S}{4}.$$

The isorent contours for quarter-CBD firms are thus as shown in figure 4*b*.

Which size of transaction area yields the highest bid rent at a given location depends on the trade-off between travel costs at that location and the gains from specialization. Not surprisingly, whole-CBD firms are viable at the center, if anywhere, and quarter-CBD firms are viable in the corners, if anywhere. The noncircular bid-rent contours of quarter-CBD firms do, however, generate a variety of equilibrium occupancy patterns, as illustrated in figure 4*c*–*g*. The changes from left to right result from decreases in NC and/or from increases in $q(F)$ relative to $q(F/4)$.

Any one of figure 4*c*–*g* suggests the same qualitative generalization to a world in which firms can vary their degree of specialization and transaction area continuously. For example, in each broad line of business, the more specialized firms will be found nearer the center. Among lines of business, those where economies of specialization are exhausted at a relatively small scale will be found primarily on the periphery.



(a) Areas in which different rent functions apply

(b) Iso-rent contours

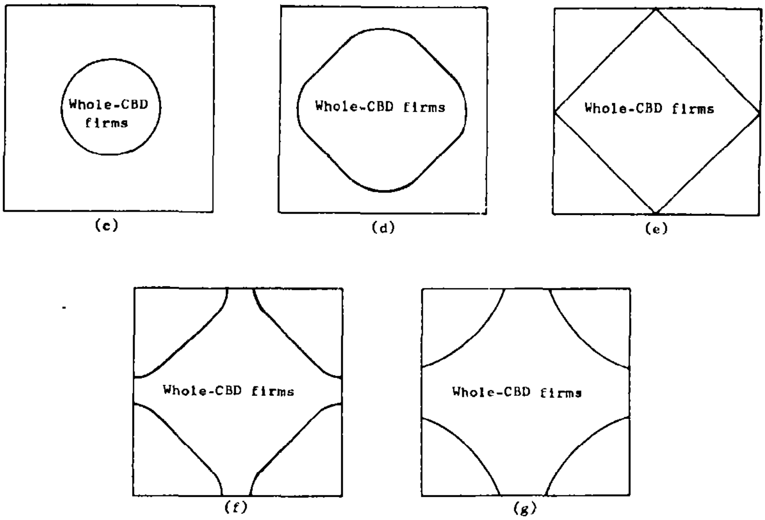


FIG. 4.—Rent contours for firms that transact only with one-quarter of the CBD, and division of the CBD between whole-CBD firms and quarter-CBD firms. The sequence from left to right in fig. 4*c*–*g* represents decreases in NC and/or increases in $q(F)$ relative to $q(F/4)$.

The whole sequence of figure 4*c*–*g* might also be interpreted dynamically as representing steps in the process by which a geographical area comes to be a unified business district composed entirely of firms doing business over the whole area. This would occur as travel costs declined or as changes in technology or growth in the surrounding metropolitan area increased the economical degree of specialization. Although it is not clear how much of figure 4*c*–*g* is produced by the artificially simple assumptions of the model, the role of the central axes is at least suggestive of real-world patterns.

VI. Concluding Remarks

This paper has investigated the implications of the face-to-face transactions that are implicit in "agglomerative economies." The resulting travel-to-all-locations model produces, in its simplest form, a concave-from-below office-rent function. A possibly fortuitous result is the dependence of rent on radial distance even though travel is rectilinear. The concavity of the office-rent function is shown to hold even for arbitrary firm distributions. For endogenous firm distributions, both the distribution of firms (or building heights) and the land-rent function are concave near the origin, but not necessarily throughout the CBD. Numerical investigation of endogenous firm distributions suggests that rent contours remain nearly circular for plausible parameter values. Indeed, the relationship between office rent and the distribution of firms, though nominally simultaneous, is nearly recursive.

Thus, more detailed analyses of other aspects of CBD production can proceed on the initial assumption of a uniform firm distribution. Two modest attempts in this direction, one involving firms with varying transactions intensity and one with firms of varying ability to benefit from agglomerative economics, suggest that the concavity of office rent will persist in more elaborate models. For example, by relaxing the assumption of equiprobable transactions, it should be possible to analyze the development of "districts" of firms among which transactions are especially likely. A related situation involves retail stores: here the customer wants to purchase from several types of stores; hence the closer each store is to other stores, the more customers it will attract. Finally, with a more detailed (and probably nonconvex) specification of the firm's production technology, it might be possible to analyze how a firm chooses its degree of specialization by balancing the direct benefits of specialization against the travel costs needed to realize agglomerative economies.

References

Alchian, Armen. "Costs and Outputs." In *The Allocation of Economic Resources*, edited by M. Abramovitz et al. Stanford, Calif.: Stanford Univ. Press, 1959.

Hoover, Edgar M. "The Evolving Form and Organization of the Metropolis." In *Issues in Urban Economics*, edited by Harvey Perloff and Lowdon Wingo, Jr. Baltimore: Johns Hopkins Press, for Resources for the Future, 1968.

Mills, Edwin S. "The Value of Urban Land." In *The Quality of the Urban Environment*, edited by Harvey Perloff. Baltimore: Resources for the Future, 1969.

Miyao, Takahiro. "A Note on Land Use in a Square City." Unpublished paper, Univ. California, Santa Barbara, July 1976.

Muth, Richard. *Cities and Housing*. Chicago: Univ. Chicago Press, 1969.

Segal, David. "Are There Returns to Scale in City Size." *Rev. Econ. and Statis.* 58 (August 1976): 339–50.

Solow, Robert M., and Vickrey, William S. "Land Use in a Long Narrow City." *J. Econ. Theory* 3 (December 1971): 430–47.

VanRensselaer, Marianna Griswold. *Henry Hobson Richardson and His Works.* Houghton-Mifflin, 1888; reprint ed. Chicago: Prairie School, 1967.

Vernon, Raymond. *Metropolis 1985.* Cambridge: Harvard Univ. Press, 1960.

Transaction Costs and Interest Arbitrage: Tranquil versus Turbulent Periods

---

Jacob A. Frenkel

University of Chicago and Tel-Aviv University

Richard M. Levich

New York University

This paper deals with the effects of transaction costs on the efficacy of covered interest arbitrage during three periods: 1962–67, the tranquil peg; 1968–69, the turbulent peg; and 1973–75, the managed float. Several conclusions emerge: (i) during the managed float transaction costs have risen dramatically, (ii) these costs played a similar role in accounting for deviations from parity during the periods of the tranquil peg and the managed float but not during the turbulent peg. Similar conclusions emerge from a time-series analysis of the various exchange rates with the implication that a classification of periods according to the degree of turbulence is preferred to a classification based on the legal arrangement (e.g., pegged or floating rates), and (iii) covered interest arbitrage does not seem to entail unexploited opportunities for profits.

This paper deals with transaction costs and the efficiency of short-term capital movements. Our prime purpose is to examine whether, and to what extent, the efficiency of short-term capital movements has been affected by different economic environments generated by the varying exchange rate regimes.

Previous versions of this paper were presented under the title "Transaction Costs and the Efficiency of International Capital Markets" at the Sixth Konstanz Seminar on Monetary Theory and Policy, Konstanz, Germany, June 1975, and at a conference on Monetary Mechanism in Open Economies, Helsinki, Finland, August 1975. This paper was completed while R. M. Levich was a visiting scholar in the Division of Finance of the Board of Governors of the Federal Reserve System. The views expressed herein are solely those of the authors and do not necessarily represent those of the Federal Reserve System. J. A. Frenkel gratefully acknowledges a Ford Foundation grant. The authors wish to acknowledge the useful comments of a referee.

[*Journal of Political Economy*, 1977, vol. 85, no. 6]

© 1977 by The University of Chicago. All rights reserved.

In a previous paper (Frenkel and Levich 1975), we suggested a method for estimating the cost of transactions that is associated with covered interest arbitrage. The period studied was January 1962–November 1967, a period which was relatively homogeneous in terms of the extent of volatility in the foreign exchange markets. We refer to that period as the "tranquil pegged exchange rate period." The British devaluation in November 1967 terminated the "tranquil period" and set the stage for extremely turbulent times in the foreign exchange markets during 1968–69, a period that may be referred to as "the turbulent pegged exchange rate period." Following a transitional phase, the system evolved into its present state referred to as the managed (or dirty) float.

In the present paper we apply the method of estimation to the above mentioned three phases of the exchange rate system and concentrate on comparing the results obtained for these three periods. One of the conclusions emerging from such a comparison is reflected in the title of this paper and relates to the classification of periods; it suggests that for the study of covered interest arbitrage, it might be preferred to classify periods by the extent of turbulence rather than by the legal and institutional arrangements of the exchange rate regime (e.g., pegged or flexible exchange rate systems). In Section I we outline some of the theoretical aspects associated with incorporating the cost of transactions into the standard formulae of covered interest arbitrage. Section II contains a description of the methodology and the estimates of the cost of transactions. In Section III we use the estimated cost of transactions to assess their significance in accounting for the observed deviations from the parity condition during the three periods under examination. In Section IV we examine the implications of following a simple trading rule. In Section V we proceed in comparing some of the characteristics of the three periods by examining the time-series properties of the various exchange rates. Some concluding remarks are contained in Section VI.

I. Interest Arbitrage and Transaction Costs: Theoretical Aspects

The interest parity theory states that the equilibrium forward premium on foreign exchange is:

$$\frac{F - S}{S} = \frac{i - i^*}{1 + i^*}, \tag{1}$$

where F and S are, respectively, the forward and spot exchange rates, and where i and i^* are the domestic and the foreign rates of interest on securities that are identical in all respects except for the currency of denomination. The formulation in equation (1) ignores any transaction costs in the security and the foreign exchange markets. In Frenkel and

Levich (1975) we have shown that, when costs are present (and are proportional to the value of transactions), the *lower* limit on the forward premium $p \equiv (F - S)/S$ for which covered *outflow* is profitable is:

$$\underline{p} = \frac{(1 + i) - \Omega(1 + i^*)}{\Omega(1 + i^*)}, \tag{2}$$

where $\Omega = (1 - t)(1 - t_S)(1 - t^*)(1 - t_F)$ and where t, t^*, t_S, and t_F denote, respectively, the percentage cost of transactions in domestic and foreign securities and in spot and forward exchange rates.[1]

By similar reasoning it can be shown that the *upper* limit on the forward premium for which covered *inflow* is profitable is:

$$\bar{p} = \frac{\Omega(1 + i) - (1 + i^*)}{1 + i^*}, \tag{3}$$

which is smaller than the value indicated from equation (1).

Equations (2) and (3) set the limits for a neutral band within which covered interest arbitrage is not profitable. Whenever the forward premium falls within the neutral band such that

$$\frac{\Omega(1 - i) - (1 + i^*)}{1 + i^*} \le p \le \frac{(1 + i) - \Omega(1 + i^*)}{\Omega(1 + i^*)}, \tag{4}$$

there will be no incentive for covered interest arbitrage.[2] Thus, points that are bounded within the neutral band may be viewed as equilibrium points even though the condition of equation (1) is not satisfied. The width of the neutral band increases with the cost of transactions. To enable such an analysis it is necessary, therefore, to estimate these costs.[3]

---

[1] To illustrate, consider the costs of covered outflow. If arbitragers do not hold cash, then covered outflow requires the execution of four transactions: (*a*) sale of domestic securities with transaction costs of t percent, (*b*) spot purchase of foreign currency with transaction costs of t_S percent, (*c*) purchase of foreign security with transaction costs of t^* percent, and (*d*) forward sale of foreign currency with transaction costs of t_F percent.

[2] *The above formulation assumes that potential arbitragers always hold securities*; as a result, the initial transaction has to be a sale of securities for cash. Since, however, at each period some fraction of the existing stock of securities matures, some arbitragers may initially hold cash. Allowing for this possibility in the above formulae results in a more conservative measure of the cost of transactions where in equation (2) we replace Ω by $\Omega_1 = \Omega/(1 - t)^2$ and in equation (3) we replace Ω by $\Omega_2 = \Omega/(1 - t^*)^2$ with the corresponding modification of condition (4). Our framework corresponds to the typical description which views arbitrage as a "round trip" activity. E.g., U.S. arbitragers begin and end with dollar denominated assets and avoid exchange risk by covering in the forward market. To the extent that holders of a specific currency wish to purchase another currency without completing the round trip, some of the costly transactions could be avoided. In this case the resulting band will be narrower than indicated by (4). For an elaboration of this case, see Deardorff (1977).

[3] Observed deviations from the arbitrage condition induced various attempts to reconcile the theoretical expectations with the empirical facts. In addition to covered interest arbitrage, the modern theory of forward exchange considers also the activities of hedging and speculation as factors determining the forward rate (e.g., Tsiang 1959;

II. The Cost of Transactions: Methodology and Estimates

There are no direct estimates of the cost of transactions in either the market for foreign exchange or the security market. In an earlier paper (Frenkel and Levich 1975), we proposed a procedure for the estimation of the cost of transactions in the market for foreign exchange by using data on triangular arbitrage. Briefly, the essence of triangular arbitrage is to ensure consistency of cross exchange rates. As an example, consider the exchange rates between the U.S. dollar and the U.K. pound, the U.S. dollar and the German mark, and between the German mark and the U.K. pound. In the absence of transaction costs, arbitrage among currencies ensures that equation (5) holds:

$$(\$/£)_\tau = (\$/\mathrm{DM})_\tau(\mathrm{DM}/£)_\tau, \tag{5}$$

where the terms in parentheses indicate the corresponding exchange rates, and the subscript τ indicates that these prices are for foreign exchange delivered at the same maturity τ.[4] If transactions in the foreign exchange market are costly, the two sides of equation (5) could differ, and the maximum discrepancy would correspond to the differential costs involved in executing the two types of exchange. Furthermore, if the cost of transacting is approximately equal across leading currencies, then one would expect the maximum discrepancy between the two sides of equation (5) to correspond to the cost of one transaction. Accordingly, this is the way that we estimate the cost of transactions in the market for foreign exchange. This estimate should be interpreted to encompass the total cost associated with a transaction. Thus it includes elements like brokerage fees, time cost, subscription costs, and all other components that comprise the cost of being informed.

Taking the cost of transactions in the market for foreign exchange as the upper limit of the discrepancy from triangular arbitrage is based on the assumption that the structure of the cost remains stable throughout the period under examination. It is only under this assumption that we

---

Kenen 1965; and Grubel 1966). Other attempts to account for the persisting deviations emphasized the role of transaction costs (Branson 1969; Frenkel and Levich 1975), political risk (Aliber 1973), the role of the elasticities of demand and supply in the capital markets (Prachowny 1970; and Frenkel 1973) as well as differential tax treatments (Levi 1977). For a survey, see Officer and Willett (1970).

[4] This formulation assumes that the various exchange rates reflect a direct exchange of one currency for another. In practice, many exchanges between two convertible currencies may use another vehicle currency as an intermediary. Since all of the exchange rates in equations (5) and (5′) involve either the U.K. pound or the U.S. dollar (both being vehicle currencies) it seems that the assumption of a direct exchange is satisfied. It should be noted that our method may not be generalized for cases which do not involve vehicle currencies since not all exchange rates will reflect meaningful direct quotations. We are indebted to an anonymous referee for raising this issue.

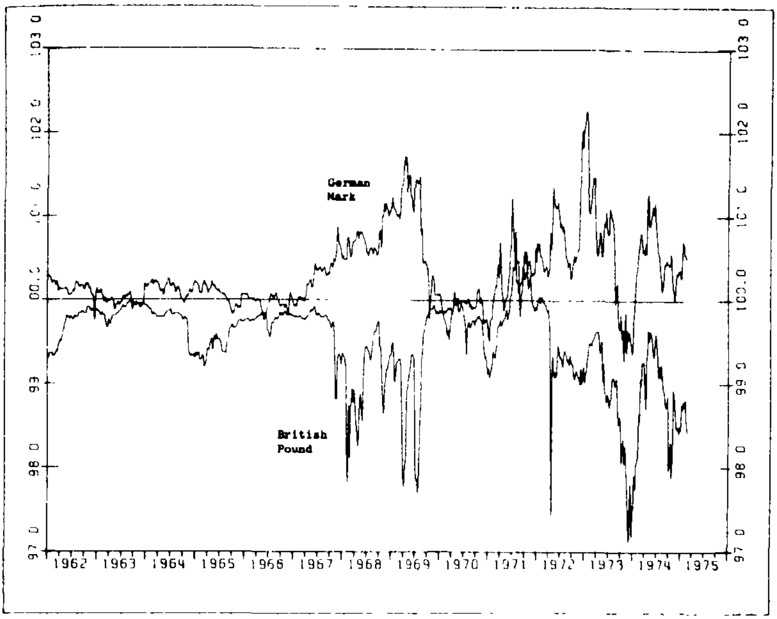


Fig. 1.—Weekly observations of the ratio of 90-day forward to spot exchange rates

can interpret smaller deviations from triangular arbitrage as being within a neutral band such that the cost of transaction exceeds arbitrage profits. Therefore, it is necessary to identify periods during which market conditions were approximately homogeneous. In identifying such periods we follow the suggestion of Leamer and Stern (1972) and inspect the ratio of the forward to the spot exchange rates for various currencies. A forward rate lying outside the support limits may indicate (under a regime of pegged exchange rates) pressures reflecting lack of confidence in the government's ability to maintain the peg. When the exchange rate is flexible, there is no obligation to maintain a specific rate; in this case periods may be classified according to the degree of volatility of the ratio of the forward to the spot exchange rate.

In figure 1 we plot the ratio of the 90-day forward to the spot exchange rates for two pairs of currencies: the mark-dollar and the pound-dollar rates. As is evident, several periods emerge: (i) 1962–67, which we identify as the tranquil peg; (ii) 1968–69, which we identify as the turbulent peg; and (iii) 1973–75, the managed float. We chose not to analyze the period January 1970 June 1973 which marks the breakdown of the pegged rate system and the transition towards the current regime.

As can be seen from figure 1 the various peaks and troughs correspond to the major crises in the international financial markets.[5] To further explore the relative homogeneity of the three periods, we have examined the deviations from triangular arbitrage; while these deviations differ markedly among periods, their structure is relatively stable within each period. Accordingly, we have concluded that there are three distinct periods for which we need to estimate the cost of transactions in the market for foreign exchange.

To estimate the cost of transactions we have analyzed the series of weekly observations of the spot and the 90-day forward exchange rates.[6] Using these observations we have computed for each period two series of weekly percentage deviations from the triangular arbitrage of equation (5) for the spot and the forward rates. The upper limits of these series of deviations correspond to t_S and t_F, respectively—the cost of transactions in the spot and the forward markets. To allow for errors in measurement and other data inaccuracies, we have taken a more conservative measure: our estimates of t_S and t_F for the various periods are the percentage deviation which bounds 95 percent of the weekly deviations from triangular arbitrage.

The indirect exchange between the U.S. dollar and the U.K. pound in equation (5) was assumed to occur through the German mark. In principle, other currencies could be used as intermediaries. It is expected that competition will assure that the cost of transacting will tend to be equalized among leading vehicle currencies. To allow, however, for possible differences in transaction costs as a function of the intermediate currency, we have also computed the cost for the case in which the Canadian dollar (\$C) replaces the German mark as the vehicle of arbitrage as in equation (5′):[7]

$$(\$/£)_\tau = (\$/\$C)_\tau(\$C/£)_\tau. \tag{5′}$$

The various estimates of the cost of transactions in the market for foreign exchange are summarized in table 1. As can be seen, these estimates differ significantly among the different periods under consideration while within each period the estimates show much less variation with respect to the choice of the intermediate currency.

The changes in the cost of transactions are also related to the pattern that is depicted in figure 1. The British devaluation in November 1967 signified the start of the turbulent period which manifested itself in lack of confidence and increased uncertainty. In the face of these developments

---

[5] A more comprehensive Crisis Index which provides for a similar classification is developed by Stokes (1972).

[6] All data sources used in this paper are outlined in the Appendix.

[7] For a related discussion on the role of the cost of transactions in the determination of vehicle currencies, see Swoboda (1969).

TABLE 1

ESTIMATES OF THE PERCENTAGE COST OF TRANSACTIONS IN THE MARKET FOR FOREIGN EXCHANGE
(Spot and 90-Day Forward)

Period and Arbitrage between Securities	Intermediate Currency	t_S	t_F	$t_S + t_F$
Jan. 1962–Nov. 1967:				
$US and £	DM	.051	.076	.127
$US and £	$C	.058	.068	.126
Jan. 1968–Dec. 1969:				
$US and £	DM	.102	.160	.262
$US and £	$C	.085	.112	.197
July 1973–May 1975:				
$US and £	DM	.523	.507	1.030
$US and £	$C	.438	.442	.920

dealers require a larger compensation for carrying inventory of various currencies and for standing ready to take immediate positions.

In the most recent managed float period, the cost of transactions in foreign exchange has risen to unprecedented heights. Depending on the currency and on the maturity of the contract, the cost of transactions during the period is between six to 10 times higher than the corresponding cost during the tranquil peg.[8]

It should be emphasized, however, that these developments need not be interpreted as a necessary characteristic of the flexible exchange rate regime. They may have been just coincidental to the timing of the system's evolution into the current regime or they may have reflected the friction associated with the transition toward and the familiarization with a new exchange rate system. There is, however, a presumption that ask-bid spreads, which are components of the cost of transactions, are higher during periods of uncertainty. Volatility of price series may imply that price changes contain new information. Dealers, wishing to protect themselves against the superior information that may be possessed by several traders, will quote a wider ask-bid spread and thereby raise the cost of transactions. Indeed, evidence on the early float indicates the close link between the extent of uncertainty and the ask-bid spread.[9] In addition,

---

[8] As we mentioned before, we define the managed float period as of July 1973 rather than March 1973 (when the U.S. dollar was devalued) to allow for a transition period

[9] Using daily observations on the ask-bid spread for various currencies during the early float in 1971, Fieleke (1975) found a close link between the spread and some proxies for the extent of uncertainty. Our approach to the estimation of the cost of transactions in the market for foreign exchange differs in that we do not interpret the size of the daily deviation from triangular arbitrage as the daily cost. Rather, we prefer to provide an estimate for a longer period that is more or less homogeneous. Further evidence of the pattern displayed by the ask-bid spread is provided in Aliber (1975), McCormick (1975), and McKinnon (1976). An interesting theoretical analysis on the relationship between uncertainty and the ask-bid spread is contained in Allen (1977).

TABLE 2

ESTIMATES OF THE PERCENTAGE COST OF TRANSACTIONS IN THE MARKETS FOR 90-DAY SECURITIES

Period	U.S. Treasury Bills—t	Euro \$ rate t (External)	Euro £ rate t^* (External)
Jan. 1962–Nov. 1967	.0095	N.A.	N.A.
Jan. 1968–Dec. 1969	.0132	.0381	.1172
July 1973–May 1975	.0299	.0381	.1175

if transaction costs include fixed components, the average cost of transaction is inversely related to the size of transactions. To the extent that the rise in uncertainty reduced the average size of transactions, it may have contributed to the higher cost.

The second class of transactions required for covered interest arbitrage involve the security markets. We adopt Demsetz's (1968) estimate according to which total costs are about 2.5 times the ask-bid spread. Since the ask-bid spread corresponds to the cost of two transactions (a round trip), our estimate of the cost of a *single* transaction is 1.25 times the ask-bid spread.

In the next section we explore the role of transaction costs on the efficacy of covered interest arbitrage among various groups of securities. One group comprises traditional pairs of securities: the U.S.-U.K. Treasury bills and the U.S.-Canadian Treasury bills. The second group comprises external pairs of securities, for example, securities that are issued in the Euro-market.[10] Since we did not have a uniform source of information on the spreads in the Euro-currency markets for the period as a whole, we report estimates only for the latter two periods. The modal values of these costs for the various periods are reported in table 2. As can be seen from table 2, the cost of transactions in the U.S. Treasury bills increased over the period under examination, while the cost in the Eurodollar market remained stable as between the last two periods. The cost in the thinnest market—the Euro-sterling market—exceeded the corresponding cost in the other markets.

III. Interest Arbitrage and Transaction Costs: Empirical Aspects for Tranquil and Turbulent Periods

In this section, we use the formulae developed in Section I together with the estimates of the cost of transactions to examine their quantitative

---

[10] Aliber (1973) has argued that the traditional pair does not always satisfy the comparability criterion since Treasury bills are issued in financial centers that may differ in political risks. In that respect the Euro-market securities are more comparable since both are issued in an external center. There is some evidence that banks use the Euro-currency interest rates in making their quotations of the forward exchange rate. To this extent the interest parity relation in the Euro-market will be extremely robust.

TABLE 3

Transaction Costs and Neutral Band for 90-Day Covered Arbitrage

Period and Arbitrage between Securities	Intermediate Currency for Estimating t_S and f_F	Observations Bounded within Neutral Band (%)			
		Traditional Pair		External Pair	
		Ω	Ω'	Ω	Ω'
Jan. 1962–Nov. 1967:					
$US and £	DM	87.0	82.4	99.7	99.3
$US and £	$C	87.0	82.1	99.7	99.3
$US and $C	$C	87.6	82.1	N.A.	N.A.
Jan. 1968–Dec. 1969:					
$US and £	DM	36.9	35.0	97.1	87.4
$US and £	$C	33.0	30.1	94.2	69.9
$US and $C	$C	67.0	63.1	N.A.	N.A.
July 1973–May 1975:					
$US and £	DM	89.7	87.6	100.0	99.0
$US and £	$C	84.5	81.4	99.0	99.0
$US and $C	$C	100.0	99.0	100.0	100.0

impact on the operation of covered interest arbitrage. Of special interest is the comparison of the efficacy of covered interest arbitrage—when allowance is made for the cost of transactions—in exploiting profit opportunities during the specified periods.

The traditional pairs are analyzed by examining weekly data on arbitrage of 90-day U.S. and U.K. Treasury bills as well as between the 90-day U.S. and Canadian bills. The external pair is analyzed by examining weekly data on arbitrage between 90-day Euro-dollar and Euro-sterling securities.[11]

The neutral band around the interest parity line was computed by using the various estimates of the cost of transactions in foreign exchange—t_S and t_F—and the modal estimates of the cost of transactions in securities—t and t^*.

The fraction of the observations bound within the neutral band generated by the cost of transactions during the various periods is reported in table 3. The computations were made for the traditional as well as the external pairs of securities using the alternative estimates of the cost from tables 1 and 2. In computing the neutral band, arbitragers were assumed to start from two alternative initial positions: first, where the initial position is in securities so that the cost corresponds to Ω, and second, where the initial position is in cash so that the cost corresponds to Ω_1 and Ω_2; the latter case is reported in table 3 under the column headed by Ω'.

---

[11] As was emphasized in Frenkel and Levich (1975), to conform to the 90-day 1-period model, we have transformed the annualized data back into their 90-day counterparts by dividing the published premia and interest rates by four. The need for such a transformation was emphasized by Goschen (1864, pp. 193–243). The following empirical analysis is based on the transformed data.

Due to lack of data on external Canadian interest rates, a similar analysis could not be applied to arbitrage between U.S. and Canadian yields in the external market.

Table 3 reveals that the cost of transactions played a very similar and significant role during the tranquil peg (1962–67) and the managed float (1973–75) periods. In both of these periods the cost of transactions accounted for a similar fraction of the apparent profit opportunities implied by deviations from the interest parity condition. This result is striking in view of the vast differences in the estimates of the cost of transactions during these two periods. Within each of these periods the proportion of the deviations from the U.S.-U.K. parity line explained by the cost of transactions is similar to the proportion of the deviations from the U.S.-Canada parity line explained by these costs. Thus, during these two periods the role played by transaction costs in accounting for deviations from the interest parity line was almost invariant with respect to the arbitraged assets. The observation that similar fractions of the deviations from parity are explained by the cost of transactions leads to a conclusion that, in spite of the large differences in the estimates of the cost, there has been no fundamental structural change concerning the relative role of the cost of transactions. This result is of some interest in view of the basic changes that took place in the international financial system as well as in the exchange rate regime between 1962–67 and 1973–75.

The intermediate period of the turbulent peg during 1968–69 seems to differ fundamentally. During that period the cost of transactions accounts for a much smaller proportion of the deviations from the parity. Moreover, the fraction of the deviations that is explained by transaction costs differs between the pairs of the arbitraged assets. However, when we consider the deviations from parity of a different pair of securities—the external pair—almost all of the deviations are accounted for. It thus seems that unlike the periods 1962–67 and 1973–75, the traditional pairs of securities were not comparable during the period 1968–69. The turbulent period of 1968–69 was characterized by financial uncertainty, a reduced cooperation among central banks, and related phenomena which comprise the concept of "political risk."[12] Much of these elements of incomparability is removed when the arbitrage is within the Eurocurrency market. What remains to be seen, however, is whether the observed deviations from parity (after allowance for transaction costs) imply some sort of market imperfections in the sense of unexploited profit opportunities; this issue is analyzed in the following section.

---

[12] Viewing the period 1968–69 as a turbulent speculative period suggests that, on theoretical grounds, the exact parity condition need not hold even when due account is given to the cost of transactions (Tsiang 1959; and Kenen 1965). In addition to the cost of transactions the width of the neutral band around the interest parity line depends on the size of the elasticities of the demand and supply in the capital markets and in the foreign exchange markets. The lower the elasticities, the wider becomes the neutral band. In Frenkel and Levich (1975) we have developed the formulae which take account of the possibility that the various elasticities are finite. Applying these formulae to the various periods revealed that, given the cost of transactions, the implied elasticities were very high (reaching at least the value of 100) even during the period 1968–69.

The comparison of the various periods also suggests that, for the study of covered interest arbitrage, periods should be classified by the extent of turbulence, speculative pressures, and the like rather than by the legal arrangement of the exchange rate regime (e g., pegged or flexible exchange rate systems). We will return to examine this implication in Section V.

IV. Unexploited Profits and a Simple Trading Rule

The previous analysis suggests that a significant fraction of the deviations from parity is accounted for by allowing for cost of transactions. Since the remaining observations (most abundant during the period 1968 69) indicate the possibility of unexploited profit opportunities, a closer examination of this phenomenon might be useful.

To explore the possibility of exploiting the apparent profit opportunities we have examined the implications of an extremely simple trading rule: investors receive information at period t, and whenever there is a possibility of a profit opportunity (after allowing for transactions cost) a transaction is executed at period $t + 1$ (at the prices prevailing at $t + 1$).[13] Table 4 summarizes the implications of such a simple trading rule for the alternative measures of the cost of transactions and for the alternative assumptions concerning the initial positions of arbitragers. If investors were able to transact in quoted prices at period t, there would be N profit opportunities. The implied mean percentage profit is indicated in the corresponding entry with a t-statistic in parentheses below the mean. For example, during the tranquil period 1962 67 the mean percentage profit would have been about 0.05 percent, while during 1973-75 it would have been about 0.3 percent. However, when actual transactions are executed at period $t + 1$, some of the profit opportunities turn out to be illusory, and the mean percentage profits are reduced to about 0.04 percent during 1962–67 and to about 0.1 percent during 1973–75. The turbulent period 1968 69 proves again to be somewhat of an exception. During that period the number of apparent profit opportunities as well as the mean rate of profit are high, and the simple trading rule still leaves relatively high and statistically significant profits. However, as argued before, during that period the traditional pairs of securities do not seem

---

[13] The lag between the receipt of information and the actual execution of arbitrage transactions may be rationalized in terms of the technology of information gathering and processing. For the distinction between ex ante and ex post profits, see Frenkel and Levich (1975).

TABLE 4

MEAN PERCENTAGE PROFITS OF A SIMPLE TRADING RULE
(Traditional Pair)

PERIOD AND ARBITRAGE BETWEEN SECURITIES	INTERMEDIATE CURRENCY FOR ESTIMATING t_S AND t_F	N		t (Mean Profit)		$t + 1$ (Mean Profit)	
		Ω	Ω′	Ω	Ω′	Ω	Ω′
Jan. 1962–Nov. 1967:							
$US and £	DM	40	54	.055	.058	.038	.045
				(7.2)	(8.4)	(3.7)	(5.4)
$US and £	$C	40	55	.056	.058	.039	.044
				(7.3)	(8.4)	(3.8)	(5.4)
$US and $C	$C	38	55	.058	.058	.046	.041
				(7.8)	(8.7)	(3.8)	(4.2)
Jan. 1968–Dec. 1969:							
$US and £	DM	65	67	.577	.585	.540	.550
				(9.6)	(9.8)	(8.3)	(8.6)
$US and £	$C	69	72	.606	.607	.572	.586
				(10.2)	(10.4)	(9.1)	(9.4)
$US and $C	$C	34	38	.107	.120	.086	.102
				(6.8)	(7.9)	(4.1)	(5.2)
July 1973–May 1975:							
$US and £	DM	10	12	.261	.272	.076	.097
				(4.3)	(4.6)	(.8)	(1.1)
$US and £	$C	15	18	.304	.306	.115	.187
				(5.1)	(5.3)	(1.2)	(2.0)
Jan. 1968–Dec. 1969 (external pair):							
$US and £	DM	3	13	.310	.154	−.319	−.210
				(4.1)	(3.2)	(6.4)	(2.7)
$US and £	$C	6	31	.193	.104	−.357	−.055
				(2.2)	(3.9)	(3.08)	(1.5)

NOTE.—t-statistic in parentheses

to be comparable, and thus they do not seem to be appropriate for the analysis of covered interest arbitrage. Indeed, when the same computation is applied to the external pair of securities (see the lower panel of table 4), the positive and significant mean percentage profit that would have been implied if transactions could be executed at t turn out to entail significant losses under the simple trading rule.[14]

V. A Time-Series Approach

One of the themes of discussion in Section III was a comparison among the various periods. The implication of that comparison was that it might

---

[14] While the above results illustrate the relevance of distinguishing between the ex-ante and ex-post concepts of profit opportunities and the possible implications of a simple trading rule, they should be viewed with caution. A conclusive test of the trading rule requires much more refined data than used in the present computation. Ideally, the difference between t and $t + 1$ should correspond to consecutive transactions or at most to daily data. Since the results of table 4 are based on weekly data, they should be viewed only as illustrative.

TABLE 5

TIME-SERIES MODELS FOR SPOT EXCHANGE IN THREE PERIODS

TIME PERIOD AND COUNTRY	MOVING AVERAGE COEFFICIENTS θ_1	θ_2	θ_3	$10^{-5}\hat{\sigma}_a^2$	Q (k)	2 ln λ (r)
1962-67 ($N = 306$):						
England	−.06 (1.0)	−.06 (1.1)	−.11 (1.9)	.021	22.8 (21)	1.46 (3)
Germany	−.10 (1.6)	−.14 (2.3)	...	.050	17.2 (22)	6.08* (2)
Canada......... .. .	−.10 (1.6)	−.17 (3.1)	. .	.334	12.4 (22)	9.92* (2)
1968-69 ($N = 99$):						
England	−.10 (.9)	−.07 (.7)	+.15 (1.5)	.167	16.2 (21)	1.18 (3)
Germany	−.10 (.9)	. .		3.420	6.9 (23)	−2.05 (1)
Canada......... .	+.14 (1.4)	−.16 (1.7)	...	.123	12.3 (22)	1.60 (2)
1973-75 ($N = 96$):						
England	−.16 (1.6)	+.10 (1.0)	−.40 (4.1)	7.480	16.3 (21)	10.22* (3)
Germany	+.10 (1.0)	−.47 (5.5)	+.18 (2.0)	20.500	18.4 (21)	16.32* (3)
Canada.....	−.10 (1.0)	−.24 (2.3)	.	.852	18.1 (22)	4.19** (2)

NOTE.—Moving average coefficients in the model ln $(S_t/S_{t-1}) = a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} - \theta_3 a_{t-3}$ with t-statistics in parentheses. $\hat{\sigma}_a{}^2$ = residual variance in series, Q = Box Pierce test statistic for autocorrelation of residuals distributed as χ^2 with k degrees of freedom, 2 ln λ = Zellner-Palm test statistic for significance of fitted model against null hypothesis that the series is a random walk—distributed as χ^2 with r degrees of freedom.

* Significant at 5% level

** Significant at 20% level

be preferable to classify periods according to the degree of turbulence rather than the legal arrangements of the exchange rate regime. In the present section we pursue further the comparison among periods by examining the time-series processes of the various exchange rates.[15]

In estimating the time-series processes we apply the Box-Jenkins time-series approach. Using the first differences of the logarithms of the various exchange rates, we computed the sample autocorrelation functions so as to identify the models, which were then estimated. The maximum likelihood parameter estimates for the two time series (the spot and the forward exchange rates) are reported in tables 5 and 6. As can be inferred from the Q-statistics, the residuals are not serially correlated (except for one case), and thus we cannot reject the hypothesis that the transformation reduced the observed data to white noise.[16]

---

[15] A more detailed analysis of the time-series models is contained in Levich (1977). For a clear introduction to the applications of the approach, see Nelson (1973).

[16] In fitting the series and deciding whether to add additional terms we have used the Zellner-Palm (1974) likelihood ratio analysis which tests objectively the hypothesis that additional terms reduce the residual variance. The Zellner-Palm test statistic is reported in the last columns of tables 5 and 6.

TABLE 6

TIME-SERIES MODELS FOR FORWARD EXCHANGE IN THREE PERIODS

TIME PERIOD AND COUNTRY	MOVING AVERAGE COEFFICIENTS			$10^{-5}\sigma_a^2$	Q (k)	2 ln λ (r)
	θ_1	θ_2	θ_3			
1962–67 (N = 306):						
England	−.08	+.02	−.25	.033	31.3*	12.60*
	(1.4)	(.4)	(4.5)		(21)	(3)
Germany	−.13	..	..	.041	20.3	4.42*
	(2.2)				(23)	(1)
Canada	−.05	−.16	..	.354	11.0	6.84*
	(.9)	(2.9)			(22)	(2)
1968–69 (N = 99):						
England	−.05	−.04	−.05	1.900	18.9	−2.64
	(.5)	(.4)	(.5)		(21)	(3)
Germany	−.07	..		3.180	14.6	−2.20
	(.7)				(23)	(1)
Canada	+.07	...	..	1.690	16.2	−0.59
	(.7)				(23)	(1)
1973–75 (N = 96):						
England	−.06	+.02	−.37	9.800	19.4	8.43*
	(.6)	(.2)	(3.8)		(21)	(3)
Germany	+.05	−.40	.	2.230	14.3	16.38*
	(.5)	(4.5)			(22)	(2)
Canada	−.10	−.22	.	.788	18.2	3.12
	(1.0)	(2.1)			(22)	(2)

NOTE.—For an explanation, see note to table 5.
* Significant at 5% level.

In comparing the general characteristics of the time-series properties of spot and forward rates, the first pattern to notice is that in the first and last periods the random walk hypothesis is rejected at the 5 percent level in nine of the 12 series (as is evident from the Zellner-Palm test). On the other hand, during the turbulent peg period (1968–69) we cannot reject the random walk model for any of the six series.[17]

A closer examination reveals further similarities between the tranquil peg and the managed float periods. Consider for example the pound spot rate: both the tranquil peg and the managed float periods contain a significant third moving-average term which is not significant in the turbulent peg period. This phenomenon is even more striking when we compare the estimates for the British pound's forward rates in table 6: during the tranquil peg the estimates of the moving-average coefficients are −.08, +.02, −.25, while during the managed float the respective coefficients are −.06, +.02, −.37, with significant third moving-average

[17] Since exchange rates reflect the outcomes of economic policies, we should emphasize that there is nothing special in the random walk hypothesis and no inference concerning efficiency may be drawn from rejecting the random walk. Our purpose in this section is more descriptive with the intent of showing that the first and the third periods are similar.

terms.[18] The similarity in the sign and magnitude of these coefficients for these two periods stands in contrast with the estimated time-series process for the turbulent peg which appears to follow a random walk.

The Canadian exchange rate also provides an example: during the tranquil peg and the managed float periods the second moving-average coefficients are significant for both the spot and the forward rates in contrast with the turbulent peg period for which the exchange rates series may be characterized as a random walk. The German-mark case is somewhat weaker although the 1962-67 and 1973-75 periods seem to have more in common with each other than with the 1968-69 period.

To sum up, the evidence from the time-series analysis seems to be consistent with the interpretation that the tranquil peg and the managed float periods are similar to each other while both differ from the turbulent peg period.

VI. Concluding Remarks

In this paper we examined the effects of transaction costs on the efficacy of covered interest arbitrage during three periods: 1962-67, the tranquil peg; 1968-69, the turbulent peg; and 1973-75, the managed float.

Several conclusions emerge; first, the cost of executing transactions associated with covered interest arbitrage has risen dramatically during the managed float period as compared with the previous periods; second, in spite of the vast differences in the estimates of the cost, they played a similar quantitative role in accounting for deviations from parity during the tranquil peg and the managed float periods. The exceptional period was that of the turbulent peg. This pattern is also consistent with evidence provided by the time-series characteristics of the various exchange rates. It suggests that for the purpose of analyzing covered interest arbitrage, it might be preferred to classify periods by the degree of turbulence rather than by the legal arrangement of the exchange rate regime.[19] The third conclusion concerns the efficacy of arbitrage in eliminating profit opportunities. The data suggest that—after allowing for transaction costs and ensuring that the arbitraged assets are comparable—covered interest arbitrage does not seem to entail unexploited opportunities for profit.

Our conclusion on the empirical validity of the interest parity relationship is of significance for the recently developed asset market approach to exchange rate determination. An important characteristic of this approach is the hypothesis that interest parity holds and that the speed of adjustment is relatively high.[20] Our paper provides evidence that is consistent with this building block of the asset approach to the determination of exchange rates.

---

[18] The appearance of what seems to be a 3-week cycle in the British pound has also been recorded by Grubel (1966) and more recently by Upson (1972) who applies a spectral analysis.

[19] For a similar classification see Stein (1962).

Appendix on Data

I. Foreign Exchange Market

Exchange rates involving the U.S. dollar and used to display the behavior of the series (fig. 1) and to calculate the cost of transactions (table 1) for the period 1962-69 are taken from the *International Financial Statistics* (IFS), International Monetary Fund (IMF), Washington. These data were collected by the IMF as follows: the \$U.S./£ rates are the closing prices in London as reported in the *Financial Times*. The \$U.S./\$C rates are the noontime interbank prices in Toronto as reported by the Bank of Canada, and the \$U.S./DM rates are the 11:00 A.M. Official Session Quotation of the Bundesbank in Frankfurt. Data on the latter period used in figure 1 are from the *Weekly Review of International Money Markets*, Harris Bank, Chicago. These data are the closing bid prices from the New York interbank market for the last trading day of the week. For the purpose of computing transaction costs during 1973–75, data involving the \$U.S. are from the *International Monetary Market Year Book* and the *Daily Information Bulletin* published by the staff of the International Monetary Market (IMM) of the Chicago Mercantile Exchange and are closing midpoints.

The exchange rates involving the U.K. pound sterling (DM/£) and (\$C/£) are from the *Montagu Monthly Review*, Samuel Montagu and Co., Ltd., London. The spot rates are the midpoints of the daily range. The forward rates were obtained by adding the midpoints of the closing forward spreads to the estimates of the spot rates (described above).

Ideally, in computing transaction costs one should allow for the fact that some transactions use the ask price while some use the bid price. Our use of the midpoints may introduce some noise but need not introduce a systematic bias. Similarly, the various exchange rates should be quotations at the same moment in time; since our data sources report prices that are few hours apart, an additional source of noise is introduced. To allow for these inaccuracies, we have used conservative measures which bound 95 percent of the deviations from triangular arbitrage.

II. Security Market

All 3-month Treasury bill rates used in the analysis of the interest parity for the traditional pairs of securities for the period 1962-69 are taken from the *Federal Reserve Bulletin*, various issues. In fact, since the *Federal Reserve Bulletin* reports the latest 3-month rate, some of the rates used are for bills that are slightly less than 3 months by 2 or 3 days. For the period 1973–75 the data source was the *Weekly Review of International Money Markets*, Harris Bank, Chicago.

The external rates used for the period 1962–69 are Euro-sterling deposits in Paris and Euro-dollar deposits in London: both are 3-month middle closing rates taken from the *Bank of England Quarterly Bulletin*, various issues. For the period 1973–75 these rates are 3-month deposits in Frankfurt taken from the *Money Manager* (New York).

---

[20] See, e.g., Frenkel (1976), Dornbusch (1976), Bilson (1978), and related papers in the *Scandinavian Journal of Economics*, vol. 78, no. 2 (May 1976).

The computation of the cost of transactions in securities (table 2) is based on the percentage ask-bid spread. The bid price and the ask price (expressed as a percentage of par) were computed according to (i) bid price = 100 – (bid yield × days to maturity)/360; (ii) ask price = 100 – (ask yield × days to maturity)/360, and the percentage spread is (ask price – bid price)/ask price. The data on the bid yield and ask yield on U.S. Treasury bills for the period 1962–69 are from *Salomon Brothers Monthly Bond Report*, while for the later period they are from the *Money Manager*. Since 90-day Treasury bills are not issued every day, for each observation date we have computed the ask-bid spread on the Treasury bill whose maturity was closest to 90 days. The range of maturities accepted was between 87 and 93 days. The ask-bid spread on the external deposits for the period 1973–75 were taken from the *Money Manager* and for the period 1968–69 from the *Bond Buyer* (which in 1972 changed its name to the *Money Manager*).

In the absence of information on ask-bid spreads on the U.K. and the Canadian Treasury bills, we have assumed that those are equal to the spread on the U.S. bills and thus that, for the traditional pairs of securities, $t = t^*$. Since transaction costs are expected to be inversely related to the width of the market, the assumption that $t = t^*$ might have introduced a downward bias in the estimate of the cost. As indicated in the discussion of table 2, we had no estimates for the period 1962–67 of the cost of transactions in Euro-dollar and Euro-sterling deposits. In tables 3–4 for that period we have used instead the cost of transactions in the 90-day U.S. Treasury bills. For the reason indicated above, this might have resulted in a conservative estimate of the cost of transactions in the external market for the early subperiod.

References

Aliber, Robert Z. "The Interest Rate Parity Theorem: A Reinterpretation." *J.P.E.* 81, no. 6 (November/December 1973): 1451–59.

———. "Monetary Interdependence under Floating Exchange Rates." *J. Finance* 30 (May 1975): 365–76.

Allen, William A. "A Note on Uncertainty, Transaction Costs and Interest Parity." *J. Monetary Econ.* 3, no. 3 (July 1977): 367–73.

Bilson, John F. O. "Rational Expectations and the Exchange Rate." In *Studies in the Economics of Exchange Rates*, edited by Jacob A. Frenkel and Harry G. Johnson. Reading, Mass.: Addison-Wesley, 1978.

Branson, William H. "The Minimum Covered Interest Differential Needed for International Arbitrage Activity." *J.P.E.* 77, no. 6 (November/December 1969): 1028–35.

Deardorff, Alan V. "Do Transactions Costs Really Account for Departures from Interest Parity." Manuscript. Univ. Michigan, 1977.

Demsetz, Harold. "The Cost of Transacting." *Q.J.E.* 82, no. 1 (February 1968): 33–53.

Dornbusch, Rudiger. "The Theory of Flexible Exchange Rate Regimes and Macroeconomic Policy." *Scandinavian J. Econ.* 78, no. 2 (May 1976): 255–75.

Fieleke, Norman S. "Exchange-Rate Flexibility and the Efficiency of the Foreign

Exchange Markets." *J. Financial and Quantitative Analysis* 10, no. 3 (September 1975): 409–28.

Frenkel, Jacob A. "Elasticities and the Interest Parity Theory." *J.P.E.* 81, no. 3 (May/June 1973): 741–47.

———. "A Monetary Approach to the Exchange Rate: Doctrinal Aspects and Empirical Evidence." *Scandinavian J. Econ.* 78, no. 2 (May 1976): 200–224.

Frenkel, Jacob A., and Levich, Richard M. "Covered Interest Arbitrage: Unexploited Profits?" *J.P.E.* 83, no. 2 (April 1975): 325–38.

Goschen, George J. *The Theory of the Foreign Exchanges.* 3d ed. London, 1864. Reprinted 4th ed., London: Pitman, 1932.

Grubel, Herbert G. *Forward Exchange, Speculation, and the International Flow of Capital.* Stanford, Calif.: Stanford Univ. Press, 1966.

Kenen, Peter B. "Trade, Speculation, and the Forward Exchange Rate." In *Trade, Growth and the Balance of Payments*, edited by R. Baldwin et al. Amsterdam: North-Holland, 1965.

Leamer, Edward E., and Stern, Robert M. "Problems in the Theory and Empirical Estimation of International Capital Movements." In *International Mobility and Movement of Capital*, edited by Fritz Machlup et al. New York: Nat. Bur. Econ. Res., 1972.

Levi, Maurice D. "Taxation and 'Abnormal' International Capital Flows." *J.P.E.* 85, no. 3 (June 1977): 635–46.

Levich, Richard M. "The International Money Market: Tests of Forecasting Models and Market Efficiency." Ph.D. dissertation, Univ. Chicago, 1977.

McCormick, Frank. "Transaction Costs in the Foreign Exchange Market under Fixed and Floating Exchange Rates." Paper presented to the Western Economic Association, San Diego, Calif., 1975.

McKinnon, Ronald I. "Floating Foreign Exchange Rates 1973–74: The Emperor's New Clothes." Carnegie-Rochester Conference Series on Public Policy. *J. Monetary Econ.*, suppl. 3 (1976): 79–114.

Nelson, Charles R. *Applied Time Series Analysis for Managerial Forecasting.* San Francisco: Holden-Day, 1973.

Officer, Lawrence H., and Willett, Thomas D. "The Covered-Arbitrage Schedule: A Critical Survey of Recent Developments." *J. Money, Credit, and Banking* 2, no. 2 (May 1970): 247–57.

Prachowny, Martin F. J. "A Note on Interest Parity and the Supply of Arbitrage Funds." *J.P.E.* 78, no. 3 (May/June 1970): 540–45.

Stein, Jerome L. *The Nature and Efficiency of the Foreign Exchange Market.* Essays in International Finance, no. 40. Princeton, N.J.: Internat. Finance Sec., Princeton Univ., 1962.

Stokes, Houston H. "The Crisis Index: An Empirical Test of the Degree of Tension in the Foreign Exchange Market." *Econ. Notes* 1, no. 1 (January–April 1972): 72–91.

Swoboda, Alexander K. "Vehicle Currencies: The Case of the Dollar." In *The International Market for Foreign Exchange*, edited by Robert Z. Aliber. New York: Praeger, 1969.

Tsiang, S. C. "The Theory of Forward Exchange and the Effects of Government Intervention on the Forward Market." *IMF Staff Papers* 7, no. 1 (April 1959): 75–106.

Upson, Roger B. "Random Walk and Forward Exchange Rates: A Spectral Analysis." *J. Financial and Quantitative Analysis* 7, no. 4 (September 1972): 1897–1905.

Zellner, Arnold, and Palm, Franz. "Time Series Analysis and Simultaneous Equation Econometric Models." *J. Econometrics* 2 (May 1974): 17–54.

Money Wages, Prices, and Causality

Y. P. Mehra

Pennsylvania State University

Using recently developed statistical techniques for examining the causal patterns between two variables within a bivariate distributed lag system, it is shown that, for the U.S. historical sample period 1954–70, money wages and consumer prices are simultaneously determined. This bidirectional feedback structure between wages and prices appears very strongly at the manufacturing level, though there is some evidence that such a structure does hold at the industry level, too. The structure of the causal patterns observed between industry money wages and prices for this sample period is not related to the industry market structure.

The main purpose of this paper is to report the results of empirical tests for determining the causal patterns between industry money wages and consumer prices. It is rather intuitive that industry money wages via their effects upon industry product prices may have significant effects upon the consumer price index, and the widespread use since the mid 1950s of the escalatory clauses in union contracts has probably also increased the sensitivity of industry money wages to the consumer price index. At a less intuitive level, if money wages depend upon expected prices which are formed rationally conditioned on the past history of wages and prices, then we will also observe significant feedbacks between industry money wages and consumer prices. These considerations therefore suggest that there are distributed lag relations between industry money wages and consumer prices, the exact shapes of these lag distributions being dependent upon the institutional characteristics of the wage-bargaining process, the market structure, the relative importance of the industry product

The author wishes to thank an anonymous referee of this *Journal* for many helpful comments on an earlier version of this paper.

[*Journal of Political Economy*, 1977, vol 85, no. 6]

© 1977 by The University of Chicago. All rights reserved.

prices in the consumption bundle, and the nature of the underlying expectation formation mechanism. Numerous published works[1] on industry wage and price determination models have recognized these interrelationships between industry money wages and prices, but no attempt has been made so far to test the structure of the causal patterns.[2] In fact there are several questions that remain to be answered. Does an increase in wage settlements achieved by a single industry's bargaining units lead to higher consumer prices? Does an increase in inflation as measured by the consumer price index lead to higher industry wage settlements? Do the causal patterns between money wages and consumer prices observed at the manufacturing level carry over to the industry level? Does the industry market structure have any role to play in the structure of causal patterns? These are the questions to which some answers are sought in this paper.

Section I describes briefly the causality tests used in this paper. The empirical results and the associated statistical inferences are reported in Section II. Section III presents some tests for serial correlation and discusses the stability of the estimated causal patterns over the subsamples. Section IV contains the concluding observations.

I. Causality Tests

Both Granger (1969) and Sims (1972) have described the statistical theory that can be used to construct tests of causal patterns within a bivariate system, and Sargent (1976) recently has used that theory to test his own formulation of the natural rate hypothesis. Sargent's work can be adapted to construct a version of Granger's causality test for the problem in hand. Consider moving average representations of W_t and CPI

$$W_i(t) = \sum_{s=0}^{\infty} a_i(s)u_i(t - s) + \delta_i LT, \qquad \sum_{s=0}^{\infty} a_i^2(s) < \infty, \qquad i = 1, \ldots, N \quad (1)$$

$$\text{CPI}\ (t) = \sum_{s=0}^{\infty} b(s)v(t - s) + \delta LT, \qquad \sum_{s=0}^{\infty} b^2(s) < \infty$$

---

[1] This literature is well documented in Nordhaus (1972) and Gordon (1972).

[2] The main exception at the aggregate manufacturing level is the recent work by Geweke (1974). Although Geweke focuses on the relationship between money wages and the wholesale price index, he does run some causality tests between the manufacturing money wage and CPI. Using monthly and quarterly data spanning the time period I 1955–I 1971, Geweke has shown that the feedback structure between manufacturing money wage and the consumer price index is bidirectional. But he did not study the causal patterns between industry money wages and the consumer prices, the main focus of the present paper.

where W_i is the ith industry money wage, CPI is the consumer price index, N is the number of industries, u's and v's are serially uncorrelated processes with mean zero and finite variances, and LT is a linear trend so that system (1) represents indeterminate covariance-stationary stochastic processes. Assume that the roots of

$$\sum_{s=0}^{\infty} a_i(s)\lambda_i^s = 0, \qquad \sum_{s=0}^{\infty} b(s)\lambda^s = 0, \qquad i = 1, \ldots, N$$

lie outside the unit circle so that system (1) possesses the autoregressive representation of the form

$$\begin{aligned} W_i(t) &= \sum_{s=1}^{\infty} g_i(s) W_i(t-s) + \delta_i LT + u_i(t) \\ \mathrm{CPI}(t) &= \sum_{s=1}^{\infty} h(s)\ \mathrm{CPI}(t-s) + \delta LT + v(t), \qquad i = 1, 2, \ldots, N. \end{aligned} \tag{2}$$

Let $Q_1(t)$ be the set of observations on CPI observed as of time t or earlier and $Q_2(t)$ be the set of observations on the relevant industry money wage observed as of time t or earlier. If, in the univariate stochastic representation of industry money wages and consumer prices (eq. [2]), the innovations $u_i(t)$ and $v(t)$ obey

$$\begin{aligned} &E[u_i(t) \mid Q_1(t-1)] = 0 \\ &E[v(t) \mid Q_2(t-1)] = 0, \qquad i = 1, 2, \ldots, N, \end{aligned} \tag{3}$$

so that the innovations in the industry money wages, $u_i(t)$, $i = 1, 2, \ldots, N$, are statistically independent of the information contained in $Q_1(t-1)$ and the innovations in the consumer price index, $v(t)$, are statistically independent of the information contained in $Q_2(t-1)$, then taking into account the information contained in the sets $Q_i(t-1)$ $(i = 1, 2)$ does not help improve the forecasts of the relevant variables by more than can be made on the basis of their own lagged values. Since this is exactly the criterion of causality proposed in Granger (1969, p. 428), one has the following practical procedure based on Granger's and Sargent's works to test the causal patterns between W_i and CPI: using the method of least squares, estimate the linear regressions of $W_i(t)$ and CPI(t) on lagged W_is and lagged CPIs as

$$\begin{aligned} \hat{W}_i(t) &= \sum_{s=1}^{m_i} \hat{a}_i(s) W_i(t-s) + \hat{\delta}_i LT + \sum_{s=1}^{n_i} \hat{e}_i(s)\ \mathrm{CPI}(t-s) \\ \mathrm{C\hat{P}I}(t) &= \sum_{s=1}^{m} \hat{b}(s)\ \mathrm{CPI}(t-s) + \hat{\delta} LT + \sum_{s=1}^{n_i} \hat{f}_i(s) W_i(t-s), \end{aligned} \tag{4}$$

where $\hat{a}_i$, $\hat{b}$, $\hat{e}_i$, $\hat{f}_i$, $\hat{\delta}_i$, $\hat{\delta}$ are the least-squares estimates. On the null hypothesis that CPI does not cause W_i, the parent parameters $e_i(s)$,

$s = 1, \ldots, n_i$, equal zero. Similarly on the null hypothesis that W_i does not cause CPI, the parent parameters $f_i(s)$, $s = 1, 2, \ldots, n_i$, should be zero. Following Sargent, I will call this Granger's test of causality.

An alternative way of testing the causal patterns as hypothesized in (4) is suggested by Sims (1972). According to Sims's test, CPI causes W_i if, in the two-sided distributed lag relation of CPI on W_i, the leading values of W_i have regression coefficients significantly different from zero as a group, whereas W_i causes CPI if, in the two-sided distributed lag relation of W_i on CPI, the leading values of CPI have regression coefficients significantly different from zero. In particular, least-squares estimation yields the following form for two-sided distributed lag relations between an ith industry's money wage and the CPI:

$$W_i(t) = \alpha_i + \sum_{s=-l_1}^{l_2} \beta_i(s)\,\mathrm{CPI}(t-s) + \varepsilon_i(t) \qquad i = 1, 2, \ldots, N \quad (5a)$$

$$\mathrm{CPI}(t) = \delta_i + \sum_{s=-l_1}^{l_2} \gamma_i(s) W_i(t-s) + \eta(t) \qquad t = 1, 2, \ldots, T \quad (5b)$$

where l_1 and l_2 are, respectively, the length of the lag distribution on the leading and lagging values of the relevant independent variables and $\varepsilon_i(t)$ and $\eta(t)$ are the error terms. On the null hypothesis that W_i does not cause CPI, $\beta_i(s) = 0$ for $s < 0$ in (5a). Similarly under the null hypothesis that CPI does not cause W_i, $\gamma_i(s) = 0$ for $s < 0$ in (5b).

In this paper both Granger's and Sims's tests are used to examine the causal patterns between industry money wages and consumer prices. The time-series on industry money wages are the straight-time average hourly earnings of the production workers at the 18 two-digit-level manufacturing industries as well as the manufacturing sector and the sample period covered is I 1954 IV 1970. The choice of the sample period is dictated by the consideration that wage price controls (war- or peacetime) may distort the true underlying relations between industry money wages and consumer prices, and hence for the investigation in hand I have chosen the period which I hope is free of such imposed distortions.

Since Sims's test requires estimating two-sided lag distributions between industry money wages and consumer prices and then running F-tests on the group significance of leading values of the independent variables, it is imperative that serious attention be paid to generating an uncorrelated error structure in (5). This is so because F-tests in general are highly sensitive to the presence of autocorrelation in the residuals. Sims (1972) used a quasi-second-differencing of the natural logarithms of the data series to reduce serial correlation in the residuals. In particular, each data series $X(t)$ was replaced by $\ln X(t) - 2k \ln X(t-1) + k^2 \ln X(t-2)$ with the value of $k = .75$. One must caution against the naive application of this particular filter to correct for autocorrelation in equations of the form (5a) and (5b). There is no reason, a priori, to expect that a linear filter which creates an uncorrelated error structure for a particular wage-price relation of the form (5a) will also generate an uncorrelated error structure for the same inversed wage-price relation of the form (5b). Hence I have employed the following empirical procedure for the choice of an appropriate value for k for use in a second-order filter of the form $1 - 2kL + k^2L^2$, $0 < k < 1$, to create an uncorrelated error structure in equations (5a) and (5b): (i) form an appropriate specification of the two-sided wage-price relations (including also linear trend and seasonal dummies), use some ad hoc filter (say, $k = \frac{3}{4}$) to prefilter the variables in that specification and apply OLS to the filtered variables to estimate the lag profile; (ii) given the least-squares residuals from step (i), evaluate the autoregressive properties of these residuals by running an autoregression on these least-squares residuals; (iii) if the autoregression indicates a purely uncorrelated autoregressive error structure, then the initial choice of the ad hoc value of k was appropriate; (iv) if the autoregression indicates the presence of some serial correlation, however, then change the value of k and repeat steps (i)-(iii). One should repeat the process until a value of k, $0 < k < 1$, is found which generates an uncorrelated autoregressive error structure as evidenced by the autoregression of the least-squares residuals.[3]

For Sims's test, one also has to specify, a priori, the length of the lag distribution on the leading and lagging values of the independent variables. In the empirical work, four leading and four lagged values were chosen in (5). For Granger's test, four own lagged terms and four lagged terms of the variable in the set Q were allowed. However, I do describe later how sensitive the results are to this particular pattern of lag truncation. Especially for Sims's test, I have considered versions of equations (5) with (8 past, 4 future) lags as well as (8 past, 8 future) lags to compare with (4 past, 4 future) lag versions of (5).

II. Empirical Results

Tables 1-6 report the results of performing causality tests along the lines proposed by Sims and Granger for the quarterly data on the 18 industry money wages and the consumer price index spanning the period I 1954-IV 1970. For Sims's test, table 1 first reports some summary statistics relating to the estimation of one-sided and two-sided unrestricted lag distributions of the wage-price and price-wage relations for the various industries. Table 2 summarizes the F-statistics pertinent for testing the

---

[3] This empirical procedure has also been used in Mehra (1976). See also Williams, Goodhart, and Gowland (1976).

TABLE 1

SUMMARY OF OLS REGRESSIONS

INDUSTRY	Eq. (A): $W = f$(CPI,4Past)			Eq. (B): $W = f$(CPI,4Past,4Future)			Eq. (C): CPI $= f$(W,4Past)			Eq. (D): CPI $= f$(W,4Past,4Future)		
	F (5,48)[a]	SEE	k	F (9,44)[a]	SEE	k	F (5,48)[a]	SEE	k	F (9,44)[a]	SEE	k
Tobacco	0.9	.0184	...[b]	0.5	.0192	...[b]	2.3*	.004005	.65	6.9*	.003013	.65
Textile	0.7	.0106	...[b]	0.7	.0108	...[b]	11.4*	.003021	.65	6.6*	.003044	.65
Apparel	7.2*	.01092	.45	10.6*	.00847	.45	10.2*	.003118	.65	6.1*	.003118	.65
Leather	14.7*	.00763	.45	17.9*	.00588	.45	15.0*	.002791	.65	9.7*	.002705	.65
Lumber	14.9*	.01103	.35	13.4*	.00951	.45	13.9*	.002854	.65	9.5*	.002722	.65
Furniture	4.4*	.00682	...[b]	3.0*	.00676	...[b]	20.7*	.002885	.60	15.2*	.002641	.60
Paper	27.2*	.00428	.35	19.0*	.00396	.35	20.6*	.003333	.55	24.2*	.002529	.55
Chemicals	12.1*	.00607	.35	10.5*	.00536	.35	8.7*	.003710	.60	15.3*	.002631	.60
Petroleum	2.0	.0116	...[b]	2.7*	.0107	...[b]	10.3*	.004108	.55	22.2*	.002622	.55
Rubber	1.0	.0092	...[b]	1.1	.0091	...[b]	6.1*	.003501	.65	7.4*	.002947	.65
Food	15.8*	.00523	.35	10.9*	.00494	.35	2.8*	.003619	.65	9.8*	.002692	.65
Fab. met.	19.4*	.00486	.45	13.1*	.00459	.45	10.1*	.004125	.55	27.2*	.002410	.55
Mach.	40.7*	.00419	.35	28.02*	.00386	.35	18.3*	.003464	.55	28.4*	.002364	.55
Elec. mach.	2.2	.00718	...[b]	2.0	.00699	...[b]	3.1*	.003886	.65	7.6*	.002917	.65
Trans.	18.24*	.00608	.35	12.07*	.00581	.35	5.6*	.003555	.65	8.8*	.002790	.65
Inst.	27.4*	.00366	.45	17.6*	.00352	.45	14.1*	.003761	.55	21.9*	.002639	.55
Stone	9.2*	.00609	...[b]	5.9*	.00535	...[b]	21.0*	.003310	.55	24.8*	.002504	.55
Pri. met.	7.9*	.01150	.35	5.9*	.01084	.35	1.8	.004103	.65	7.3*	.002963	.65
Manu.	49.8*	.00383	.35	30.15*	.00371	.35	17.3*	.002671	.65	14.07*	.002371	.65

NOTE.—All regressions on logs of the variables prefiltered by the linear filter of the form $(1 - kL)^2$, $0 < k < 1$, L lag operator, were fitted to the sample period I 1954–IV 1970. All regressions always included a constant, three seasonal dummies (even though data were seasonally adjusted), and a linear trend The appropriate value of k for each regression, found by the empirical procedure as described in the text, is indicated above for every industry regression, W is straight-time average hourly earnings of production workers in an industry and CPI is the consumer price index.

[a] Degrees of freedom

[b] For all these industry regressions, a direct estimation of serial correlation (as described in n. 12 of the text) was made and a general fourth-order autoregressive correction was done for these industries. For these, F's are distributed as $F(5,46)$, $F(9,42)$.

* Significant at 5% and the F's are a test of the null hypothesis that all right-hand-side variables except constant, linear trend, and seasonal dummies have zero coefficients.

TABLE 2

F-STATISTICS ON FUTURE COEFFICIENTS

Industry	$F1$ (4,44)^{a}	$F2$ (4,44)^{a}	$F3$ (4,36)^{a}	$F4$ (4,36)^{a}	$F5$ (8,28)^{a}	$F6$ (8,28)^{a}
Tobacco	0.10^{b}	10.21^{**}	2.61^{c**}	9.22^{**}	2.73^{d**}	17.09^{**}
Textile	0.69^{b}	0.82	0.70^{c}	1.60	0.68^{d}	0.40
Apparel	8.96^{**}	1.00	4.72^{**}	2.68^{**}	2.52^{**}	2.30^{**}
Leather	9.20^{**}	1.78	4.49^{**}	4.33^{**}	3.19^{**}	3.86^{**}
Lumber	5.14^{**}	2.20	1.46	4.67^{**}	1.03	3.84^{**}
Furniture	1.21^{b}	3.30^{**}	0.54^{c}	5.09^{**}	0.87^{d}	2.28^{**}
Paper	3.03^{**}	9.83^{**}	0.58^{c}	6.55^{**}	1.56	5.28^{**}
Chemicals	3.51^{**}	12.84^{**}	2.31^{c*}	7.66^{**}	1.59	5.22^{**}
Petroleum	3.05^{b**}	18.46^{**}	3.15^{c**}	21.91^{**}	2.50^{d**}	14.33^{**}
Rubber	1.24^{b}	5.93^{**}	1.41^{c}	7.82^{**}	2.00^{d*}	6.79^{**}
Food	2.42^{*}	10.68^{**}	0.81	8.83^{**}	1.28	5.42^{**}
Fab. met.	2.44^{*}	24.15^{**}	0.79	10.51^{**}	1.29	4.15^{**}
Mach.	3.13^{**}	14.76^{**}	0.46	4.07^{**}	0.43	2.55^{**}
Elec. mach.	1.62^{b}	10.3^{**}	1.24^{c}	11.91^{**}	1.75^{d}	9.43^{**}
Trans.	2.16^{*}	8.5^{**}	0.53	8.06^{**}	0.95	5.46^{**}
Inst.	2.02^{*}	13.37^{**}	0.45	6.26^{**}	0.89	2.53^{**}
Stone	4.39^{b**}	9.96^{**}	1.19^{c}	7.05^{**}	1.99^{d}	12.25^{**}
Prim. met.	2.40^{*}	12.02^{**}	1.16	17.22^{**}	1.62	10.98^{**}
Manu.	1.74	4.23^{**}	3.18^{**}	4.10^{**}	1.52	4.15^{**}

NOTE.—Given eq. (B) (table 1), $F1$ is a test of the null hypothesis that the four future coefficients are zero as a group and $F2$ is a test of [illegible] in eq. (D) (table 1); $F3$ is a test of the [illegible] e index are zero in $W = f$(CPI, 8 Past, [illegible] s on wages are zero in a relation CPI $= f(W$, 8 Past, 4 Future); $F5$ is a test of the null hypothesis that the eight futures on CPI are zero in $W = f$(CPI, 8 Past, 8 Future), whereas $F6$ is for the futures on wages in the reversed CPI $= f(W$, 8 Past, 8 Future). All regressions always included a constant, three seasonal dummies, and a linear trend. For each regression, a search for appropriate k in the filter $(1 - kL)^2$ was made to create uncorrelated error structure (see also text section on serial correlation) except the ones noted above where a direct estimate of the serial correlation was made and a fourth-order autocorrection was made. All regressions cover the sample period 1954–70.

[a] Degrees of freedom.

[b] These F's are distributed as (4,42).

[c] These F's are distributed as (4,34).

[d] These F's are distributed as (8,26). The equations underlying these F's use fourth-order autocorrection.

* Significant at 10% level.

** Significant at 5% level.

null hypothesis that the coefficients on the future values of the right-hand-side variable are zero, that is, the null hypothesis that the left-hand-side variable does not cause the right-hand-side variable (F's in table 2). Tables 3–5 summarize the shapes of the estimated lag profiles for all the industries as well as the manufacturing sector. Table 6 reports the F-tests pertinent for Granger causality tests on industry wages and consumer prices, that is, F-tests whether $\hat{e}_i(s)$ and $\hat{f}_i(s)$ in (4) are zero. These statistical results are consistent with the following main conclusions.

First, with respect to the issue of the causal influences from industry wages to consumer prices, we may note that, in the distributed lag relation of industry money wages on consumer prices, the four future values of consumer prices are not significant for 10 out of 18 industries at the conventional 5 percent levels of significance (see $F1$ in table 2). As one allows for a little more generous lag distribution on the past and the future values of the consumer price, the F's on the future values of the consumer

TABLE 3

ESTIMATED LAG PROFILES
$W = f(\text{CPI},4\text{Past},4\text{Future})$

INDUSTRY	LAG									STANDARD ERROR OF COEFFICIENTS	
	-4	-3	-2	-1	0	1	2	3	4	Smallest	Largest
Tobacco	0.54	-0.20	0.24	-0.67	0.68	0.32	0.41	-0.16	-0.19	0.95	1.34
Textile	-0.68	0.27	0.45	0.07	-0.03	0.53	-0.22	-0.65	-0.19	0.52	0.50
Apparel	1.2	0.43	-0.45	0.70	-0.30	-0.10	-0.94	0.22	0.20	0.41	0.51
Leather	0.18	0.79	0.20	0.34	-0.49	-0.16	0.22	-0.17	0.14	0.29	0.35
Lumber	0.87	-0.16	-0.27	1.12	-0.35	-0.45	-0.12	0.40	0.31	0.46	0.60
Furniture	-0.02	0.54	0.08	-0.05	0.12	0.92	-0.12	-0.84	0.14	0.34	0.47
Paper	-0.46	-0.09	0.35	0.40	0.16	0.33	0.007	0.22	-0.08	0.19	0.25
Chemicals	-0.74	0.07	0.11	0.43	0.51	-0.21	0.20	0.69	-0.16	0.26	0.34
Petroleum	-0.15	0.20	1.59	-0.30	0.04	-0.50	-0.02	-0.04	0.59	0.54	0.67
Rubber	-0.27	0.39	0.55	-0.02	-0.31	0.50	-0.02	0.12	-0.13	0.47	0.60
Food	-0.36	-0.32	0.45	0.15	-0.10	0.40	0.56	-0.05	0.16	0.24	0.31
Machinery	-0.30	-0.19	0.28	-0.05	0.56	0.27	-0.09	0.57	0.07	0.19	0.25
Elec. mach.	0.84	-0.09	0.17	-0.19	-0.41	1.03	-0.44	0.10	0.31	0.37	0.46
Trans.	-0.23	-0.05	-0.46	0.41	0.35	0.50	-0.13	0.44	0.37	0.28	0.37
Inst.	-0.05	-0.24	-0.19	0.15	0.48	0.32	-0.06	0.35	0.35	0.17	0.21
Stone	0.20	0.26	0.51	-0.03	0.16	0.07	0.16	-0.73	0.51	0.27	0.33
Prim. met.	-0.32	-0.56	0.35	-0.64	0.13	1.15	1.0	-0.92	1.30	0.52	0.69
Manuf.	-0.12	0.16	-0.05	0.48	-0.12	0.37	0.08	-0.08	0.27	0.18	0.24

NOTE.—These are the unrestricted generalized least-squares estimates of the two-sided lag distributions of the industry wage-consumer price relations (eqq. B in table 1).

TABLE 4

ESTIMATED LAG PROFILES
$CPI = f(W, 4Past, 4Future)$

INDUSTRY	LAG									STANDARD ERROR OF COEFFICIENTS	
	−4	−3	−2	−1	0	1	2	3	4	Smallest	Largest
Tobacco	0.05	0.10	0.12	0.14	0.13	0.09	0.08	0.08	0.03	0.02	0.03
Textile	0.02	−0.03	−0.005	0.08	0.10	0.16	0.18	0.13	0.05	0.05	0.06
Apparel	0.06	0.08	0.02	0.03	0.02	0.10	0.07	0.14	0.12	0.04	0.05
Leather	0.06	−0.03	0.12	0.02	0.01	0.18	0.06	0.22	0.12	0.05	0.06
Lumber	0.05	0.06	0.06	0.09	0.07	0.14	0.07	0.07	0.06	0.03	0.04
Furniture	−0.05	−0.06	0.06	0.22	0.16	0.09	0.11	0.14	0.09	0.05	0.07
Paper	*0.16*	*0.18*	*0.15*	*0.26*	*0.19*	*0.17*	*0.18*	*−0.02*	*−0.09*	*0.09*	*0.12*
Chemicals	0.09	0.28	0.35	0.25	0.29	0.09	−0.08	−0.12	−0.11	0.07	0.08
Petroleum	0.13	0.15	0.16	0.12	0.16	0.10	0.13	0.03	0.04	0.03	0.04
Rubber	0.08	0.17	0.18	0.17	0.11	0.12	0.17	0.12	0.003	0.04	0.06
Food	0.17	0.11	0.44	0.25	0.16	−0.01	0.01	−0.08	0.05	0.07	0.10
Fab. met.	0.20	0.27	0.16	0.24	0.09	0.03	−0.11	0.06	−0.08	0.06	0.07
Machinery	0.14	0.29	0.22	0.22	0.21	−0.13	−0.02	−0.07	0.08	0.07	0.11
Elec. mach.	0.11	0.20	0.21	0.22	0.09	0.05	0.02	0.07	0.009	0.06	0.07
Trans.	0.09	0.23	0.19	0.31	0.13	0.14	−0.09	0.03	−0.04	0.07	0.08
Inst.	0.25	0.27	0.10	0.21	0.27	−0.01	−0.09	−0.003	−0.04	0.09	0.10
Stone	0.16	0.05	0.18	0.19	0.15	0.16	0.09	0.11	−0.02	0.06	0.07
Prim. met.	0.18	0.12	0.21	0.14	0.12	0.03	0.02	−0.002	0.14	0.03	0.04
Manuf.	0.17	0.09	0.09	0.21	0.06	0.26	0.02	0.18	0.13	0.08	0.09

NOTE.—These are the unrestricted generalized least-squares estimates of the two-sided lag distributions of the consumer price-wage relations (eqq D in table 1)

TABLE 5

ESTIMATED LAG PROFILES FOR THE MANUFACTURING SECTOR

Lag in Quarters	Eq. (1): *W* on CPI	Eq. (2): *W* on CPI with Future	Eq. (3): CPI on *W*	Eq. (4): CPI on *W* with Future
−4	...	−.005	...	.170
−3	...	.273	...	.115
−2	...	.048	...	.117
−1	...	.252	...	.244
0	.176	−.346	.051	−.046
1	.672	.433	.446	.203
2	.114	.083	.366	.043
3	−.197	−.095	.410	.235
4	.347	.271	.240	.116
5	−.178	−.091	.024	.029
6	−.281	−.186	−.177	−.040
7	.090	.218	−.183	.004
8	.415	.279	−.195	−.116
SEE/100.0	.327	.297	269	.235
		Standard Error of Coefficients		
Largest	.215	.223	.111	.123
Smallest	.166	.180	.093	.092
k	35	.35	.50	.50

NOTE.—These are the estimated lag profiles of two-sided price wage relations for the manufacturing sector with (8 Past, 4 Future) versions. All regressions included, in addition to a constant, three seasonal dummies and a linear trend. Seasonal dummies were insignificant (the data used were seasonally adjusted) but trend mostly significant.

prices remain statistically insignificant at the conventional level for the majority of the industries (see *F*3 and *F*5 in table 2). On the basis of these insignificant *F*'s one is tempted to conclude that a change in an industry money wage does not cause changes in the consumer prices. However, in the case of Sims's test, the absolute size of the coefficients on the future values of the independent variable is also quite important.[4] A careful look at the estimated lag profiles reported in table 3 clearly reveals some very large coefficients on the future values of CPI in many industry wage-price relations. This pattern does not change even if we allow a more generous length of the lag distribution on past and future values of CPI.[5] This does not allow us to accept strongly the null hypothesis that there are no feedbacks from industry money wages to consumer prices.[6]

---

[4] As Sims (1972, p. 545) writes, "One should bear in mind that the absolute size of the coefficients is important regardless of the *F* value. . . If the estimated coefficients on future values are as large or larger than those on past values, bidirectional causality may be very important in practice, despite insignificant *F*'s."

[5] The estimated lag profiles associated with *F*3 and *F*5 in table 2 are not presented to conserve space.

[6] A look at the estimated lag profiles (table 3) of wage-price relations in many industries clearly indicates that the shape of the lag distribution on future as well as on past values of CPI is not well determined. This indicates that there remains severe multicollinearity

TABLE 6

GRANGER'S TEST OF CAUSALITY

Industry	*FA* (4,51)[a]	*FB* (4,51)[a]
Tobacco	0.39	7.0**
Textile	6.2**	1.2
Apparel	5.5**	0.4
Leather	16.9[b]**	0.9
Lumber	9.7[b]**	0.5
Furniture	7.7[b]**	6.6**
Paper	3.5**	3.9**
Chemicals	2.1*	8.6**
Petroleum	2.4*	2.9**
Rubber	2.2*	4.7**
Food	2.0*	7.2**
Fab. met.	3.7**	8.2**
Machinery	2.5*	5.1**
Elec. mach.	1.2	4.9**
Transport.	1.8	4.4**
Inst.	1.3	5.1**
Stone	4.2**	7.6**
Prime. met.	1.5	4.9**
Manu.	5.6**	5.5**

NOTE.—*FA* is a test of the null [illegible] (A) given below are zero and *FB* is a [illegible] 4 in equations like (B) given below [illegible] *LT* is linear trend. All regressions which included three seasonal dummies in addition to a constant were fitted to the period 1954–70. The underlying equations are

$$\mathrm{CPI}(t) = \sum_{s=1}^{4} b_s \mathrm{CPI}(t-s) + \delta LT + \sum_{s=1}^{4} f_i(s) W_i(t-s), \qquad i = 1, 2, \ldots N \tag{A}$$

$$W_i(t) = \sum_{s=1}^{4} a_{is} W_i(t-s) + \delta_i LT + \sum_{s=1}^{4} e_{is} \mathrm{CPI}(t-s). \tag{B}$$

[a] Degrees of freedom.

[b] These *F*'s are distributed as (2,55) since for these industries four own lagged values of CPI created multicollinearity. Hence we considered two own lagged values and two lagged values of the causal variable in eq. (A).

* Significant at 10%.

** Significant at 5%.

A similar weaker conclusion about the presence of feedbacks from industry money wages to consumer prices is also indicated by the results of Granger's causality test reported in table 6. For example, *FA* in table 6 is a test of the null hypothesis that the lagged values of the industry money wages do not improve the forecast of the CPI by more than can be made on the basis of the lagged CPIs, and we fail to reject the null hypothesis for 10 out of 18 industries. For these industries, we accept at 5 percent the hypothesis that industry wage changes do not cause changes in the

---

in the wage-price relations in spite of the correction for serial correlation. One may use this argument to explain insignificant *F*'s and larger coefficients on future values and conclude that there are in fact significant feedbacks from wages to CPI not detected due to multicollinearity in the CPI series. The problem with this argument is that it itself reduces one's confidence in the absolute size of the coefficients and hence in any inference derived from looking at the size of the coefficients only. This follows from the fact that, even though the estimators are unbiased in the presence of multicollinearity, for a given sample the estimated coefficients may be far from the true mean value.

consumer prices in Granger's sense. At a lower confidence level, the list is reduced to five industries, that is, the null hypothesis of no feedbacks from industry money wages to CPI is accepted at the 10 percent significance level for only five industries.[7]

Second, we may note that in the distributed lag relation of the consumer price index on the industry money wages, the four future values of the industry money wages are significant as a group (see *F*2 in table 2) for the majority of the industries. There appear some larger coefficients on the future values of the industry money wages in the various price-wage relations (see estimated lag profiles in table 4). These *F*-values as well as the size of the distributed lag coefficients on the future values of industry money wages clearly imply that the interperiod lag distributed coefficients of the industry money wages are dominated by a consistent pattern of influence from the consumer price index and hence are consistent with the inference that the consumer price index causes changes in the industry wages.

It is also clear from table 2 (see *F*4, *F*6) that the above conclusion is not sensitive to the truncation of the lag distribution at the four quarters. In fact three industries (apparel, leather, and lumber) which do not seem to indicate feedbacks from the consumer price index to industry money wages do show such feedbacks if we allow a little more generous lag distribution on the past and the future values of the money wages (compare *F*2 with *F*4 and *F*6 for these industries in table 2).

A similar conclusion is also indicated by Granger's test. In table 6 *FB* is a test of the null hypothesis that the lagged values of the consumer price index do not improve the forecast of the industry money wages by more than can be made on the basis of the lagged industry wages alone. For the majority of the industries, we can never accept the null hypothesis and this is consistent with the inference that an increase in the consumer price index causes changes in the industry money wages by Granger's test, too.

Third, the previous two inferences then imply that one cannot accept very strongly the existence of unidirectional causal ordering from consumer prices to industry money wages, an assumption implicit in most of the industry money wage determination models. There is some evidence that there are feedbacks in both directions between industry money wages and consumer prices.

---

[7] In implementing Granger's test, I specified four own lagged terms upon which a variable was permitted to depend and four lagged terms of the causal variable in the CPI-wage relations. Attempts to introduce more than four own lagged terms failed, as severe multicollinearity occurred, and it was not possible to get any estimates. So we don't know how sensitive our results are to specifying only four own lagged terms. The necessity to check the results by introducing more than four own lagged terms arises from the possibility that the variable in question may possess an autoregression of infinite order. This could result if the variable follows a mixed moving-average autoregressive stochastic process. This kind of misspecification could result in an erroneous rejection of the hypothesis of exogeneity. For more on this see Feige and Pearce (1976) and Sargent (1976).

Fourth, in contrast to rather weak evidence in favor of a bidirectional feedback structure at the industry level, the evidence is conclusive that the feedback structure at the aggregate manufacturing level is bidirectional. This is the inference consistent with both Sims's and Granger's causality tests. For Sims's test, the F-values presented in table 2 and the estimated lag profiles presented in tables 3–5 clearly indicate a feedback structure which is bidirectional. The F's presented for Granger's test in table 6 also indicate the same conclusion.

The above empirical results on the causal structure between industry money wages and consumer prices raise some interesting questions. First, is the causal structure observed at the industry level in any way related to the industry market structure? Second, does the causal structure derived from the reduced form relations of industry money wages help identify some "structures" consistent with these empirical findings?

The answer to the first question seems to be no. The finding that a rise in the consumer price index leads to higher subsequent rates of industry money wages is true for two-digit level industries irrespective of their market structure. Similarly, the weak tendency to show feedbacks from industry money wages to consumer prices seems to hold for industries that cover a wide spectrum of market structure.[8]

The answer to the second question is yes. If one assumes that the participants in the labor market take nominal prices as given and the wage relation reflects the fact that wages cannot adjust to labor market conditions instantaneously, one can write a simple version of a catch-up expectational wage-adjustment model[9] as

$$W_i(t) = k_i(t) \sum_{s=0}^{N} f_i(s)\, {}_{t-s}\hat{\mathrm{CPI}}(t) + v_i(t), \tag{6}$$

where ${}_{t-s}\hat{\mathrm{CPI}}(t)$ is the price expected at time $(t - s)$ to prevail at time t, $k_i(t)$ is some function reflecting deterministic movements in the ith industry's money wages, and $v_i(t)$ is the structural disturbance in the ith industry's wage relation. Assume that $\hat{\mathrm{CPI}}$ is formed rationally conditioned on the knowledge of past CPI's and W_i's with

$$\hat{\mathrm{CPI}}(t) = \sum_{s=0}^{\infty} C_i(s) W_i(t - s) + \sum_{s=0}^{\infty} d(s)\, \mathrm{CPI}(t - s) + u(t). \tag{7}$$

---

[8] If the length of the lag distribution is kept too short (4 past, 4 future), then there appears some tendency for the competitive industries (apparel, leather, and lumber) to show unidirectional feedbacks from wages to consumer prices. However, when the length of the lag distribution is large, then this tendency disappears in favor of a strong bidirectional feedback structure (see F's in table 2).

[9] For more on expectation hypotheses, see Lucas and Rapping (1969), Turnovsky and Wachter (1972), and Geweke (1974).

If $C_i(s) \neq 0$, then a reduced form for the industry money wages associated with (6) and (7) will in general show feedbacks in both directions between wages and prices. If $C_i(s) = 0$, then the reduced form for industry money wages will reflect a simple ordering from prices to wages, and if $d(s) = 0$, then industry money wages are exogenous.[10] The fact that both Sims's and Granger's tests on causality show strong bidirectional feedbacks between money wages and consumer prices at the aggregate level allows one to rule out versions of the structures which make either prices exogenous or money wages exogenous in reduced form relations associated with the aggregate money wage. There is evidence that such an expectational mechanism is operative at the industry level, too. The fact that we cannot strongly reject the presence of feedbacks from industry money wages to the consumer price index implies that bidirectional feedbacks may be quite important in practice.[11]

III. Tests for Serial Correlation and Subsample Stability

It is now well understood that Sims's test is highly sensitive (*a*) to the presence of serial correlation in the residuals and (*b*) to the pattern of the truncation of the lag distribution. Premature truncation of the lag distribution or the improper use of a filter to remove serial correlation may lead to erroneous rejection or acceptance of the exogeneity. I have already presented the F-statistics which allow longer lags on the past and future (table 2 longer than four quarters) and the statistical inferences discussed in this paper are based on the general pattern which emerges from all of these experiments.

As regards the serial correlation, I have always used the empirical procedure described in Section I for an appropriate k, $0 < k < 1$, for use in the second-order filter, $(1 - kL)^2$, employed to create uncorrelated autoregressive structure in reduced form relations of the form

$$W'_i(t) = \sum_{s=-4,-8}^{4,8} \beta_i(s)\,\mathrm{CPI}(t - s) + \text{trend} + \text{seasonal dummies} \tag{8a}$$

---

[10] For more on this see Sargent and Wallace (1973) and Geweke (1974).

[11] The expectational money wage determination structure is *not* the only structure consistent with the empirical results on the causal patterns reported above. Assume that the consumer prices in the wage equation simply capture "the value-added price effects." With given costs, a rise in the value added prices will raise profits and the value of the marginal product of labor to the firm. In that case, we will observe that a rise in the consumer price index will lead to subsequent higher industry money wages, for both the competitive as well as the noncompetitive sectors. On the other hand, a rise in industry money wages will raise the short-run labor costs which will be reflected in higher industry product prices and hence in higher consumer prices. Straszheim and Straszheim's (1976) recent conclusion that industry product prices of competitive industries are affected as often by short-run labor costs and demand factors as those of highly competitive industries implies that the presence of feedbacks from industry money wages to consumer prices may hold for industries regardless of their market structure.

$$\mathrm{CPI}(t) = \sum_{s=-4,-8}^{4,8} \gamma_i(s) W_i(t-s) + \text{trend} + \text{seasonal dummies} \qquad (8b)$$

as these are always the underlying relations under the alternative hypothesis for the various *F*'s reported in table 2. Presented in table 7 are the estimates of the second-order autoregression of least-squares residuals of equations like (8) reported in table 1.[12] They are consistent with the inference that prefiltering as done in this paper has created an uncorrelated autoregressive error structure[13] and therefore *F*'s presented in Section II are accurate.

To the extent that a structure described by (6) and (7) is associated with reduced form relations of the form (5), it is possible that expectations incorporate the effects of policy variables and be sensitive to the public's perception of how these variables are manipulated. These considerations suggest that there may be behavioral shifts associated with changes in economic policy—shifts which might be apparent in the consumer price-wage relation (Geweke 1974, pp. 44–45).

In order to examine this possibility, I split the sample period into I 1954–IV 1962 and I 1963–IV 1970, and implemented the Chow test to test the stability of the two-sided lag distribution of consumer prices on money wages. Except for one industry, we were never able to reject the null hypothesis that the two-sided lag distribution of consumer prices on industry money wages (including the manufacturing) was the same throughout the sample period. However, for the majority of the industries, the filter which induced an uncorrelated error structure for the entire sample never created an uncorrelated error structure for the subsamples. This suggests the possibility that there are behavioral shifts in the price-wage relation not captured by Chow's procedure. Hence I examined the causal patterns using only Sims's test for the two subsamples for the various industries.

---

[12] For some industries, the above procedure did not generate an uncorrelated autoregressive error structure. Hence for these industries, a direct estimation of the serial correlation was made and a fourth-order autoregressive correction was done (see table 1 for such industries). The procedure used is as follows: (i) form the appropriate specification as in (8), use an ad hoc filter of the form $(1 - \frac{3}{4}L)^2$ and apply OLS to filtered data to estimate the lag profile; (ii) given the estimates of the lag profile from step (i), apply those estimated coefficients to the unfiltered data to calculate the residuals; (iii) run an autoregression on these residuals and use the coefficients in this autoregression to prefilter the data in the appropriate specification and estimate the lag profile by OLS. For the asymptotic justification of these procedures, see Amemiya (1973) and Sims (1973).

[13] The only exceptions seem to be textiles, petroleum, and stone which in spite of the fourth-order autoregressive correction still indicate some negative serial correlation and hence to that extent *F*'s should be carefully interpreted. To check whether the autocorrelation may be of an order greater than two for industries that used second-order correction, I ran autoregressions on least-squares residuals of the fourth order. The estimated coefficients indicated that except for the industries noted in table 1, there was no need for an autoregressive correction of an order greater than two.

TABLE 7

Estimates of the Autoregression on Least-Squares Residuals

	Industry																		
Lag	Tob	Text	App	Lth	Lumb.	Furn.	Paper	Chem.	Pet.	Rub	Food	Fb M.	Mach	E.M.	Tran.	Inst	Stone	P M.	Manu.
Equation B $\overline{W} = f(\overline{CPI}$, 4 Past, 4 Future)																			
$t-1$	-05	-34	.07	007	-.09	-02	-.07	.01	-19	-01	21	11	05	-.16	.09	05	-44	07	.12
	(.15)	(15)	(13)	(.13)	(13)	(16)	(13)	(13)	(14)	(15)	(.14)	(13)	(13)	(.15)	(13)	(13)	(.15)	(.13)	(.13)
$t-2$	003	-.58	-01	09	-23	-.02	32	-10	-20	.05	-00	-.18	14	07	-.24	.20	-.36	-20	-.11
	(15)	(.16)	(14)	(13)	(13)	(16)	(13)	(13)	(14)	(15)	(13)	(13)	(13)	(14)	(14)	(.13)	(.16)	(13)	(.13)
$t-3$	-09	-.25				06			-08	-01		..		-00	.		-21	.	.
	(.15)	(16)				(16)			(16)	(15)				(14)			(.16)		
$t-4$	004	-05				-.15			-40	-12				-03	.		-15	.	.
	(15)	(15)				(16)			(.17)	(15)				(14)			(.15)		
Equation D. $\overline{CPI} = f(\overline{W}$, 4 Past, 4 Future)																			
$t-1$	-.04	-.005	11	03	-07	-.06	-.11	-.05	-08	04	-03	-09	-.007	-03	-.03	-03	-.03	.06	-16
	(.13)	(.14)	(.14)	(14)	(13)	(.13)	(14)	(14)	(14)	(13)	(13)	(14)	(.14)	(.14)	(.13)	(.13)	(13)	(13)	(.14)
$t-2$	.09	.16	.24	.02	.07	10	02	11	03	16	.16	12	09	09	08	.18	11	.31	-.05
	(.13)	(14)	(.14)	(14)	(13)	(14)	(.14)	(.14)	(13)	(14)	(13)	(14)	(14)	(13)	(.14)	(.14)	(13)	(14)	(.14)

Note.—These are the estimates of the second-order autoregression of least-squares residuals from eqq (B) and (D) as reported in table 1. Bar over the variables W and CPI indicates that these variables were filtered with the appropriate filters $(1 - kL)^2$, value of k as indicated in table 1 for various regressions. Standard error of the coefficients is given in parentheses.

The general pattern which emerges from these subsample causality tests is the following: (*a*) For the earlier sample period, there is some evidence that lag distribution at the manufacturing level is one-sided, with the causal ordering from wages to consumer prices.[14] However, at the industry level, there emerges evidence in favor of weak feedbacks in both directions. (*b*) For the later sample period, there is evidence in favor of weak feedbacks in both the directions. Hence, we can conclude that except for the manufacturing sector, the feedback structure that emerges from the subsamples at the industry level is in line with that emerging for the entire sample.

IV. Concluding Remarks

Using tests of causal patterns developed by Sims and Sargent, it is shown that money wages and consumer prices are simultaneously determined. This bidirectional feedback structure between money wages and consumer prices appears very strongly at the aggregate manufacturing level. There is a possibility that there are substantial bidirectional feedbacks between industry money wages and consumer prices, too, even though the statistical evidence is not very strong. It is also clear that the nature of these feedbacks is not in any way related to industry market structure. An increase in the consumer price index leads to subsequent higher money wages for both the competitive as well as the noncompetitive industries and the tendency to show weak feedbacks from industry money wages to consumer prices holds equally well for most of the industries covering a wide spectrum of market structure.

References

Amemiya, Takeshi. "Generalized Least Squares with an Estimated Autocovariance Matrix." *Econometrica* 41 (July 1973): 723–32.

Cargill, Thomas F. "An Empirical Investigation of the Wage-Lag Hypothesis." *A.E.R.* 59 (December 1969): 806–16.

Feige, Edgar L., and Pearce, Douglas K. "Economically Rational Expectations: Are Innovations in the Rate of Inflation Independent of Innovations in Measures of Monetary and Fiscal Policy?" *J.P.E.* 84, no. 3 (June 1976): 499–522.

Geweke, J. "Wage Price Dynamics in U.S. Manufacturing." Mimeographed. Univ. Minnesota, 1974.

Gordon, Robert J. "Wage-Price Controls and the Shifting Phillips Curve." *Brookings Papers on Econ. Activity* 2 (1972): 385–421.

Granger, C. W. J. "Investigating Causal Relations by Econometric Models and Cross-Spectral Methods." *Econometrica* 37 (July 1969): 424–38.

---

[14] Geweke (1974, pp. 45–46) also finds evidence favoring one-sided lag distribution with causal ordering running from wages to prices at the manufacturing level for the period I 1953–IV 1964, and this one-sidedness deteriorates as sample size is extended to include observations in the later sample period.

Hause, John C. "Spectral Analysis and the Detection of Lead-Lag Relations." *A.E.R.* 61 (March 1971): 213–17.

Lucas, R. E., and Rapping, L. A. "Price Expectations and the Phillips Curve." *A.E.R.* 59 (June 1969): 342–50.

Mehra, Y. P. "Is Money Exogenous in Money Demand Equations." Illinois State Univ., 1976.

Nordhaus, William D. "Recent Developments in Price Dynamics." In *The Econometrics of Price Determination*, edited by O. Eckstein et al. Washington: Board of Governors of the Federal Reserve System and Soc. Sci. Res. Council, 1972.

Sargent, Thomas J. "A Classical Macroeconometric Model for the United States." *J.P.E.* 84, no. 2 (April 1976): 207–37.

Sargent, Thomas J., and Wallace, N. "Rational Expectations and Dynamics of Hyperinflation." *Internat. Econ. Rev.* 17 (June 1973): 328–50.

Sims, Christopher A. "Money, Income and Causality." *A.E.R.* 62 (September 1972): 540–52.

———. "Distributed Lags." Discussion Paper no. 28, Univ. Minnesota, March 1973.

Straszheim, Donald H., and Straszheim, Mahlon R. "An Econometric Analysis of the Determination of Prices in Manufacturing Industries." *Rev. Econ. and Statis.* 58 (May 1976): 191–201.

Turnovsky, S. J., and Wachter, M. L. "A Test of Expectations Hypothesis Using Directly Observed Wage and Price Expectations." *Rev. Econ. and Statis.* 54 (February 1972): 47–54.

Williams, David; Goodhart, C. A. E.; and Gowland, D. H. "Money, Income and Causality: The U.K. Experience." *A.E.R.* 66 (June 1976): 417–23.

Exchange Rate Dynamics with Stock/Flow Interaction

Jürg Niehans

University of Bern

The paper examines the adjustment of the exchange rate and domestic prices to a monetary change under flexible exchange rates. The emphasis is on the dynamic interaction between the classical stock aspects, stressed by the monetary approach, and the flow aspects, stressed by the elasticity approach. *It is shown that even in a very simple model of* lagged asset adjustment the interaction between trade and capital flows can produce a large variety of adjustment paths, with and without overshooting, monotonic and cyclical.

Introduction

In neoclassical monetary theory, with relative prices predetermined by resources, technology, and tastes, the price level was a question of the supply and demand for money. If, in addition, it was assumed that supply and demand for money, at least in a first approximation, are independent of relative prices, the price level became a "purely monetary phenomenon." However, this simplification, while perhaps tolerably valid for equilibrium situations, could not be used during the adjustment process. During such periods, shifts in relative prices would call forth all those short-run effects on asset stocks and output flows which became known as the business cycle. While "monetarists" from Hume to Friedman stressed the asset aspects of the problem, their opponents from Malthus to Keynes emphasized output flows, but the fundamental problem was always the interaction between stocks and flows during the dynamic adjustment process.

The doctrinal history of general monetary theory is, naturally, reflected

The author gratefully acknowledges valuable comments from Trent Bertrand, Francesco Caramazza, Rudiger Dornbusch, and Jacob Frenkel.

[*Journal of Political Economy*, 1977, vol. 85, no. 6]

© 1977 by The University of Chicago. All rights reserved

in the history of international monetary analysis.[1] The classical tradition from Hume to Cassel tended to explain the exchange rate, the relative price of two currencies, in terms of the relative supply and demand for assets, particularly foreign and domestic money. This tradition was reinforced by the recognition, emerging from the debate about the transfer problem, that shifts in the relative prices of traded goods are not a crucial element in the adjustment process. Others, however, concentrated on the flows of goods and capital, which the exchange rate equilibrates in any period of time, and the price and income elasticities of these flows. The last few years have seen a much-needed revival of the monetary tradition in foreign exchange theory. One of its features, characteristically, is the tendency to downplay, by suitable simplifications, the flow aspects of the problem, just as the elasticity theorists often neglected the stock aspects. Here again, however, the central problem is still the interaction problem to which this paper is addressed.

The subject is the dynamic adjustment of the exchange rate and the domestic price level to an increase in the domestic money stock under flexible exchange rates. Recent contributions to this problem were made by Kouri (1976), Dornbusch (1976*b*) and Calvo and Rodriguez (1977). Their models postulate instantaneous asset adjustment. In the first two models this effectively eliminates trade flows from the adjustment process, in Dornbusch's model because trade is always in balance, in Kouri's model because trade flows adjust passively to the required stock adjustment. Calvo and Rodriguez, by introducing trade flows depending on relative prices, made a significant step beyond the pure stock approach, but their analysis is limited to perfect foresight and still assumes instantaneous stock adjustment. The present note focuses on the interaction between money stocks and trade flows on the assumption that asset adjustment takes time.[2] It will be shown that the menu of possible adjustment patterns, which in Dornbusch's case is limited to instantaneous overshooting, becomes much richer and more interesting. The analysis will be restricted to the simplest possible case of stock/flow interaction, abstracting from aspects which, though presumably important in reality, are not crucial from the present point of view.

The Model

Consider a "small" economy with a variable price level P trading with the rest of the world with a given foreign price level P^F. The domestic economy holds two assets, namely, domestic money M_1 and foreign money M_2, both bearing no interest. The domestic money supply is exogenously determined by the government, and while it is subject to once-over changes there is no continuing monetary expansion and thus no continuing inflation.[3] Under full employment, the domestic demand for each asset, in real terms, is assumed to be a constant-elasticity function of the ratio of the actual exchange rate E to the expected exchange rate $\bar{E}$.[4] Domestic and foreign balances are imperfect substitutes inasmuch as in full equilibrium (that is, with $E = \bar{E}$) transactions motives generally produce positive demands for each, while expected capital gains or losses result in shifts in the desired portfolio composition. The possibility of negative foreign assets, where domestic money is held by the rest of the world, is disregarded.

---

[1] On the latter see the two recent papers by Jacob Frenkel (1976*a*, 1976*b*).

[2] This is the approach Kouri put on the agenda for future research (1976, p. 283).

If logarithms are denoted by lower-case letters, the two demand functions can be written

$$m_1^* - p = a_1 + \lambda_1(e - \bar{e}), \qquad \lambda_1 > 0, \tag{1}$$

$$m_2^* - p^F = a_2 + \lambda_2(e - \bar{e}), \qquad \lambda_2 < 0. \tag{2}$$

If the current rate rises relative to the expected rate, portfolios will be shifted from foreign currency to domestic currency. Problems of expectation formation are sidestepped by the radical assumption that, for any level of m_1^*, $\bar{e}$ is equal to the corresponding equilibrium rate and that the speed with which the actual rate is expected to approach the equilibrium rate does not matter. This clearly leaves much room for greater realism.

The adjustment of actual assets to desired assets is not instantaneous, the real cash balances being changed by a constant proportion of the difference between desired and actual balances,

$$\dot{m}_1 - \dot{p} = \mu_1(m_1^* - m_1), \qquad \mu_1 > 0, \tag{3}$$

$$\dot{m}_2 - \dot{p}^F = \mu_2(m_2^* - m_2), \qquad \mu_2 > 0, \tag{4}$$

where $\dot{x} = d(\log X)/dt$. Since the domestic money supply is exogenously given and the world price level is fixed, $\dot{m}_1 = \dot{p}^F = 0$. Substituting (1) into (3) and (2) into (4), we obtain differential equations for the price level and foreign balances,

$$\dot{p} = -\mu_1[a_1 + \lambda_1(e - \bar{e}) + p - m_1], \tag{5}$$

$$\dot{m}_2 = \mu_2[a_2 + \lambda_2(e - \bar{e}) + p^F - m_2]. \tag{6}$$

---

[3] The recent literature has paid considerable attention to the possibility that as a result of a domestic monetary expansion the exchange rate may temporarily "overshoot" its new equilibrium value (Niehans 1975; Dornbusch 1976*a*, 1976*b*; Kouri 1976; Calvo and Rodriguez 1977). This possibility arises because parity of asset yields requires that a reduced level of domestic interest rates (and/or a higher level of domestic inflation) be neutralized by an expected future appreciation of the exchange rate. The present model, since it contains no interest-bearing assets (and no continuing inflation), precludes the analysis of this phenomenon, important as it is.

[4] Real foreign currency balances are measured in terms of *foreign* goods. In general, the demand for real cash balances also depends on relative prices P/P^F, but this influence is here disregarded.

These equations mean that the change in domestic prices is governed by the excess demand of domestic cash balances, while capital exports are governed by the excess demand for foreign cash balances.

Under floating rates, the change in private foreign assets is equal to the trade balance. The latter, expressed as a percentage of foreign assets, is assumed to be a linear function of the logarithm of the foreign-currency price of domestic goods, P/E:[5]

$$\dot{m}_2 = \frac{T}{M_2} = b + \beta(p - e), \qquad \beta < 0. \tag{7}$$

For short periods, β might conceivably be positive, indicating that the Marshall/Lerner condition is not satisfied. The dynamics resulting from the gradual adjustment of β from positive short-run values to negative long-term values are an important topic for future research.[6] In the present context, a constant negative value of β will be assumed.

The model consisting of (5), (6), and (7) determines the paths of p and m_2 from given initial conditions for given values of m_1, p^F, and $\bar{e}$, and it also determines the associated path of e. The characteristics of these adjustment paths are the subject of the following sections.

Comparative Statics: Full Equilibrium

The first step is to determine the properties of the full equilibrium and of its displacement. The equilibrium solution of the system is obtained by letting $\dot{p} = \dot{m}_2 = e - \bar{e} = 0$. This results in three equations in the equilibrium values of p, m_2, and e,

$$p^* = m_1 - a_1, \tag{8}$$

$$m_2^* = a_2 + p^F, \tag{9}$$

$$e^* = (b/\beta) + m_1 - a_1. \tag{10}$$

If the domestic money supply is increased, the equilibrium price level and the equilibrium exchange rate shift in proportion, while the equilibrium holdings of foreign cash balances remain the same. In terms of p and m_2, the displacement of the equilibrium is shown in figure 1 as a vertical shift from A to Q. In terms of p and e, the displacement is reflected in figure 2 as a shift from A to Q along a 45° line. Comparative statics is ruled by the quantity theory of prices and purchasing power parity.

---

[5] The constant term is $b = -\beta(p_0 - e_0)$, where $p_0 - e_0$ is the logarithm of that value of P/E which keeps the trade balance at zero.

[6] The beginning was made by Britton (1970). See also Niehans (1975), Dornbusch (1976*a*).

Trade and Capital Flows: A Preliminary View of Exchange Dynamics

The question is how the economy travels from A to Q. What I ultimately want to determine, in particular, is the dynamic adjustment path of the exchange rate. Equations (5)–(7) do not make the law of motion of the exchange rate explicit. However, some preliminary insight can be gained by considering trade and capital flows.[7] Combining (6) and (7), the condition for payments equilibrium can be written

$$b + \beta(p - e) = \mu_2[a_2 + \lambda_2(e - \bar{e}) + p^F - m_2], \tag{11}$$

which must be satisfied at all times. Equation (11) can be solved for e,

$$e = \frac{\beta}{\beta + \mu_2\lambda_2} p + \frac{\mu_2}{\beta + \mu_2\lambda_2} m_2 + c_1, \tag{12}$$

where $c_1 = (b - \mu_2 a_2 - \mu_2 p^F + \mu_2\lambda_2\bar{e})/(\beta + \mu_2\lambda_2)$. Since $\bar{e}$ is proportional to m_1 and appears in c_1 with a positive coefficient, it is evident that an increase in the domestic money supply results in an *instantaneous* depreciation, even before domestic prices and foreign currency holdings have had time to react.

Using (12) to substitute for e in (5) and (6), the changes in domestic prices and foreign assets can be expressed by two differential equations in p and m_2,

$$\dot{p} = -\frac{\mu_1}{\beta + \mu_2\lambda_2}[(\beta + \beta\lambda_1 + \mu_2\lambda_2)p + \mu_2\lambda_1 m_2] + c_2, \tag{13}$$

where $c_2 = \mu_1[\lambda_1(c_1 + \bar{e}) + m_1]$, and

$$\dot{m}_2 = \frac{\mu_2}{\beta + \mu_2\lambda_2}(p - \beta m_2) + c_3, \tag{14}$$

where $c_3 = \mu_2[\lambda_2(c_1 - \bar{e}) + p^F]$. It can easily be ascertained that in p, m_2-space the *price-equilibrium* line, defined by $\dot{p} = 0$, is rising, while the *trade-equilibrium* line, defined by $\dot{m}_2 = 0$, is falling. The two lines intersect, of course, in the point of full equilibrium. We thus obtain a phase diagram like figure 1.

Starting from the old equilibrium A, the system approaches the new equilibrium Q with a price rise which can be either monotonic or cyclical. It is inevitable that foreign assets first increase, followed by a later reduction, possibly of a cyclical nature. This means that in this model monetary expansion necessarily results in initial trade surpluses followed by deficits. This permits an intuitive preview of the implied adjustment in the exchange rate. The initial export surplus is the result of the depreciation of

---

[7] This was suggested to me by Rudiger Dornbusch.

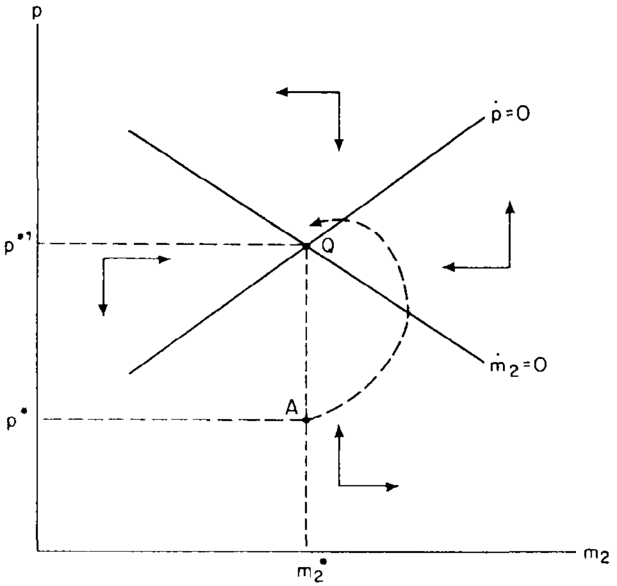


FIG. 1

the currency. However, this initial depreciation falls short of equilibrium depreciation. The reason for this "undershooting" is that the trade surplus must have as its counterpart an accumulation of foreign currency, which can only take place if more depreciation is expected in the future. Initial overshooting of the ultimate equilibrium exchange rate is thus excluded. There is nothing in the model, however, to prevent overshooting in the sense that the instantaneous depreciation is immediately followed by a temporary appreciation before further depreciation sets in. It is also possible that the exchange rate overshoots its eventual equilibrium in the later stages of a cyclical adjustment process. While overshooting is excluded in one sense of the word, it is not in others.

In terms of p and e, and thus of figure 2, the trade-equilibrium line is the 45° line going through A. Below this line there is an export surplus and foreign balances increase, while above the line there is an import surplus and foreign balances are drawn down. As was pointed out in the preceding section, the new equilibrium Q, since it is at the same level of m_2 as A, is on the trade-equilibrium line. It follows that the adjustment path cannot be wholly on one side of the trade-equilibrium line. If it does not run along the line, it must be partly below it and partly above it. If the instantaneous reaction is an increase in e, the adjustment path begins below the trade-equilibrium line and thus with surpluses, followed

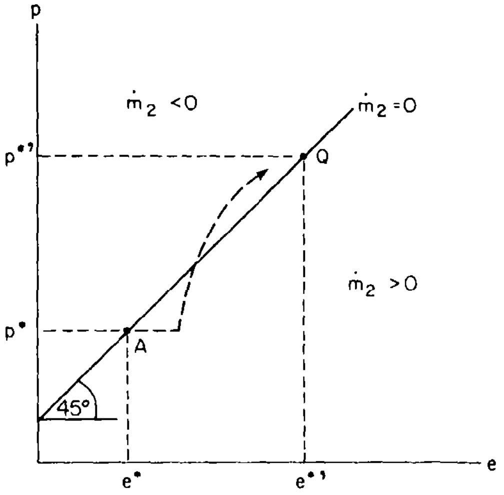


FIG. 2

by another stage above the line with deficits, as shown in figure 2 by the broken arrow. This implies that prices cannot always lag behind exchange rates or vice versa. If in one phase prices are lagging behind, in another phase they must move ahead of the exchange rate adjustment. This is what introduces relative prices into the process and makes the stock/flow interaction interesting.

Exchange Dynamics: The Phase Diagram

These intuitive arguments about the behavior of the exchange rate must now be made analytically explicit. Following a change in m_1 the exchange rate adjusts in two stages. The first stage is the *instantaneous* change necessary to keep the foreign payments in balance before foreign assets had any time to change. This reaction is triggered by the jump in the equilibrium exchange rate e^* which, by virtue of the above assumptions, is also the expected rate $\bar{e}$. We have to find those combinations of p and e which at the *new* level of $\bar{e}$, but at the inherited level of m_2, provide equilibrium in the balance of payments. The equation of this *payments-equilibrium* line, expressed in figure 3 by FF, is (12). Of course, it goes through the new equilibrium Q, and it can easily be shown that it is steeper than 45 degrees.[8] The instantaneous jump will thus be from A

---

[8] The FF-line shifts in a parallel fashion as m_2 changes in the course of the adjustment process. Trent Bertrand pointed out to me that it is tangent to the adjustment path where the latter crosses the trade-equilibrium line.

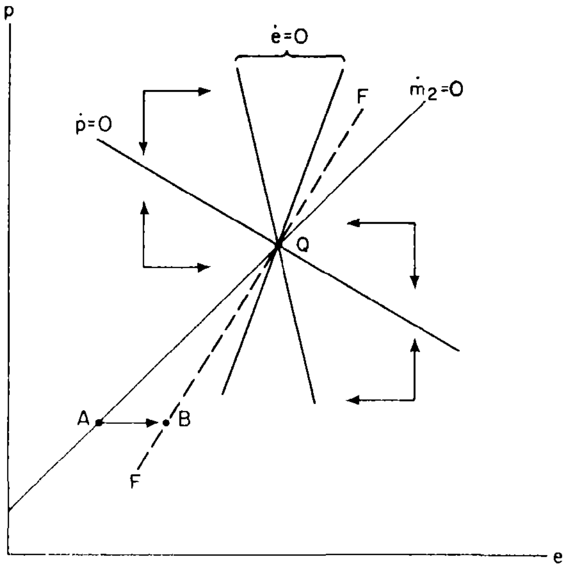


FIG. 3

to a point on the *FF*-line. The steeper this line, the larger the jump.[9] It is plausible to assume that the instantaneous adjustment occurs entirely in the exchange rate. The instantaneous reaction will thus move the system from *A* to *B*.[10] It should be noted that this timeless jump of the system is entirely due to the assumption of a timeless increase in the money supply and, as a consequence, of the expected exchange rate. With a smooth revision of expectations it would disappear It is thus a property of the analytical technique rather than of reality.

The second stage of the process is the *gradual* adjustment starting from *B*. It is governed by the differential equations of the model. The implied differential equation for e is obtained by differentiating (12) with respect to time (with $\dot{\bar{e}} = \dot{p}^F = 0$) and substituting for $\dot{p}$ from (5) and for $\dot{m}_2$ from (7). It reads

---

[9] The underlying determinants as implied in (12) can be intuitively explained as follows. While the export surplus is positively related to current depreciation and thus to e, the capital outflow is positively related to the expected future depreciation and thus negatively to e. At given p, e must be such that the export surplus just matches the capital outflow. But the export surplus depends positively on (the absolute value of) β, while the capital outflow depends positively on the adjustment speed μ_2. It follows that the instantaneous depreciation is larger the larger μ_2 and the smaller (the absolute value of) β.

[10] Under the reverse assumption that the instantaneous adjustment is entirely concentrated in prices, p would have to *fall* at first in order to produce the export surplus required to balance the capital exports resulting from the large expected depreciation of the exchange rate. This is not a plausible case.

$$\dot{e} = \frac{\beta}{\beta + \mu_2\lambda_2}\left[(\mu_2 - \mu_1)p - (\mu_2 + \mu_1\lambda_1)e + \mu_1(\lambda_1\bar{e} + m_1)\right] + c_3, \tag{15}$$

where c_3 is again a constant. For given $\bar{e}$ and m_1, (5) and (15) are a system of linear differential equations in p and e. It can be verified that with the postulated signs of the parameters the system is stable, but a "perverse" reaction of the trade balance (that is, $\beta > 0$), if strong enough, could result in instability.

To construct the phase diagram in terms of p and e, we have to determine the lines of price equilibrium and of exchange equilibrium. The *price equilibrium* line is found by letting $\dot{p} = 0$ in (5) and solving for p,

$$p = -\lambda_1(e - \bar{e}) + m_1, \tag{16}$$

which, for given m_1, has a negative slope depending on the exchange rate elasticity of the demand for money. Below this curve prices are rising, while above they are falling (see fig. 3).

The *exchange-equilibrium* line is obtained by letting $\dot{e} = 0$ in (15) and again solving for p,

$$p = \frac{\mu_2 + \mu_1\lambda_1}{\mu_2 - \mu_1}e - \frac{\mu_1}{\mu_2 - \mu_1}(\lambda_1\bar{e} + m_1) + c_4. \tag{17}$$

Since $\mu_2 \gtrless \mu_1$, this line can slope upward or downward. It will be positively sloped if the stock of foreign currency adjusts more rapidly than domestic cash balances;[11] it will be negatively sloped if the adjustment speed is higher for domestic balances. In the latter case, with $\mu_2 - \mu_1 < 0$, what can be said about the slope of the exchange-equilibrium curve relative to the slope of the price-equilibrium curve? The $\dot{e} = 0$ curve will be steeper (i.e., have algebraically a lower slope) if $(\mu_2 + \mu_1\lambda_1)/(\mu_2 - \mu_1) < -\lambda_1$ or $(1 + \lambda_1)\mu_2 > 0$. Since this inequality is always satisfied, the exchange-equilibrium line, if negatively sloped, must be steeper than the price-equilibrium line. If the adjustment speeds are the same for both assets, the exchange-equilibrium line is vertical. Nothing definite can be said about the slope of this line relative to the payments-equilibrium line; it can be either steeper or flatter. To the left of the exchange-equilibrium line, the exchange rate is rising, to the right it is falling. We thus obtain the set of possibilities illustrated in figure 3.

---

[11] *In this case it is also steeper than the trade-equilibrium curve with its 45° slope.*

The Adjustment Path

We are finally in a position to trace the adjustment path of prices and of the exchange rate under alternative conditions. In the first case, $\mu_2 < \mu_1$. This may be called the case of *slow foreign adjustment* and is illustrated in figure 4. After the instantaneous depreciation from A to B, the exchange rate continues to rise; there is no overshooting. Prices also increase, at first at an even faster rate. After a period of trade surpluses with growing foreign currency holdings, prices overtake the exchange depreciation; trade deficits appear and the extra foreign balances are gradually liquidated again. The exchange rate will approach its new equilibrium monotonically, but prices may overshoot and thus be declining in the final phase.

In the second case, $\mu_2 > \mu_1$; foreign balances are now characterized by relatively *fast adjustment*. At the same time, the exchange-equilibrium line is still assumed to be steeper than the FF-line (see fig. 5). The new element is that now the exchange rate may follow an oscillatory adjustment path in its later stages, with prices and exchange rates sometimes moving together and sometimes in opposite directions, while commodity trade alternates between surplus (capital export) and deficit (capital import). In contrast to Dornbusch's dynamic model, the overshooting of the exchange rate is not simply due to the instantaneous reaction to a discrete money supply change, but to the dynamic adjustment path itself.

In the third case, foreign balances are still assumed to adjust relatively rapidly ($\mu_2 > \mu_1$), but the exchange-equilibrium line is now less steep than the FF-line (see fig. 6). Under these conditions, the instantaneous depreciation of the domestic currency is followed by a period of gradual appreciation; the instantaneous adjustment seems to have resulted in temporary undervaluation. In due course, however, depreciation will resume and proceed either monotonically or with oscillations as described in the preceding case. In the end it will turn out that the instantaneous reaction did not, in fact, overshoot the final equilibrium, and that it was rather the temporary reappreciation which was out of line with the overall trend.

Concluding Comment

A simple stock/flow model is thus able to generate, depending on the relative size of five parameters, a rich variety of dynamic adjustment paths. This provides a strong contrast to the stark simplicity of the adjustment in pure stock models. It is highly suggestive, however, of the patterns actually observed during the last few years, where the long-term course of exchange rates seemed to conform closely to the relative rates of monetary expansion, while the short-run fluctuations showed a bewildering variety of seemingly erratic patterns.

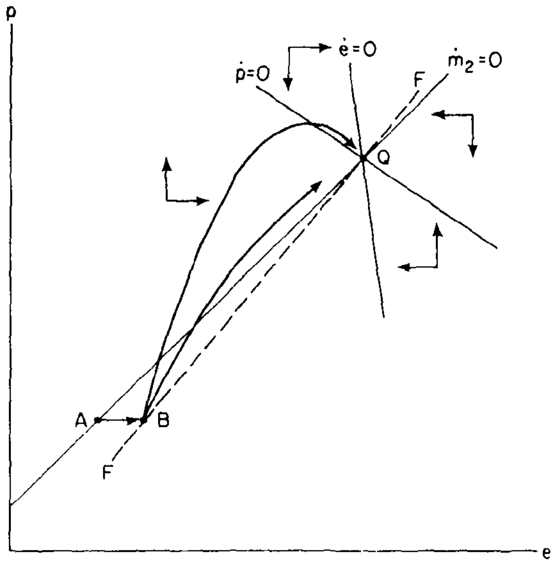


Fig. 4

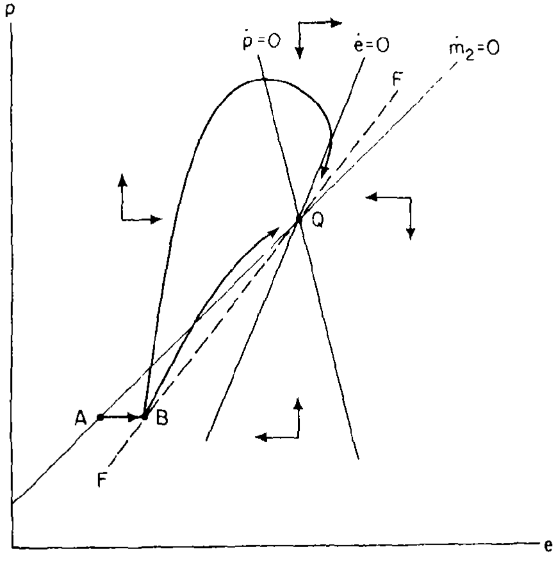


Fig. 5

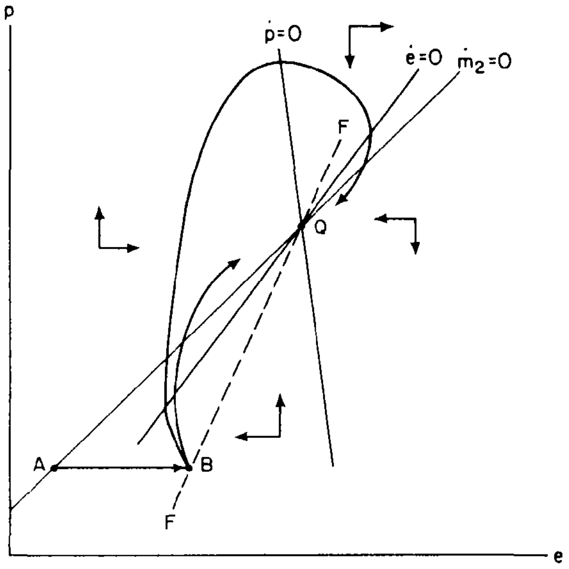


FIG. 6

It should furthermore be remembered that the present model abstracts from many important aspects of reality. The most serious deficiency, from the point of view of exchange dynamics, may be the assumption that export and import values react instantaneously to changes in prices and the exchange rate, while, in reality, this reaction is distributed over a period of years. It is quite likely that the short-term reaction is very low or even perverse because traders will be neither willing nor able to react to fluctuations which they regard as temporary. The long-term reaction may also be weak, because in the long run purchasing power parity takes over. In the intermediate run, however, the adjustment process may be significantly influenced by variations in trade flows. Another deficiency of the present model is the rudimentary treatment of expectations. A third limitation is the neglect of interest-bearing assets. Once future research will have filled these gaps, the adjustment process is likely to look still more complex with seemingly small variations in specifications and parameters producing large differences in the adjustment path. While exchange rates in the long run indeed seem to be dominated by monetary factors, it may be a long time before the adjustment pattern ceases to appear as unpredictable.

References

Britton, A. J. C. "The Dynamic Stability of the Foreign-Exchange Market." *Econ. J.* 80 (March 1970): 91–96.

Calvo, Guillermo A., and Rodriguez, Carlos Alfredo. "A Model of Exchange Rate Determination under Currency Substitution and Rational Expectations." *J.P.E.* 85, no. 3 (June 1977): 617–25.

Dornbusch, Rudiger. "Exchange Rate Expectations and Monetary Policy." *J. Internat. Econ.* 6, no. 3 (August 1976): 231–44. (*a*)

———. "Expectations and Exchange Rate Dynamics." *J.P.E.* 84, no. 6 (December 1976): 1161–76. (*b*)

Frenkel, Jacob. "Adjustment Mechanisms and the Monetary Approach to the Balance of Payments: A Doctrinal Perspective." In *Recent Issues in International Monetary Economics*, edited by E. M. Classen and P. Salin. Amsterdam: North-Holland, 1976. (*a*)

———. "A Monetary Approach to the Exchange Rate: Doctrinal Aspects and Empirical Evidence." *Scandinavian J. Econ.* 78, no. 2 (1976): 200–224. (*b*)

Kouri, Pentti J. K. "The Exchange Rate and the Balance of Payments in the Short Run and in the Long Run: A Monetary Approach." *Scandinavian J. Econ.* 78, no. 2 (1976): 280–304.

Niehans, Jürg. "Some Doubts about the Efficacy of Monetary Policy under Flexible Exchange Rates." *J. Internat. Econ.* 3 (August 1975): 275–81.

Income Distributions in Two Experimental Economies

Raymond C. Battalio, John H. Kagel, and Morgan O. Reynolds

Texas A&M University

Data on individual labor earnings are reported from two experimental economies where the primary factors responsible for income differences were differences in tastes for market income versus leisure and differences in abilities working manual job tasks. Measured income dispersion under these conditions was strikingly similar to that in the United States and other market economies, indicating that these two factors alone are sufficient to generate such income differences. Further, in tests of the functional form of the distributions of income, the hypothesis of lognormality fit better than the hypothesis of normality, just as it does in national data.

> Differences in incomes that are associated with differences in effort are generally regarded as fair. If everyone were offered the same hourly wage rate and the opportunity to work as many hours as he or she chose, the resulting discrepancies on payday would be understandable. In fact, it would seem unfair for the person who takes more leisure to get just as much income. [ARTHUR OKUN 1975]

Judging by recent literature it appears that professional interest in income distribution and inequality has rarely been higher. A key question continues to be why reported income distributions have their apparent similarity in general shape, dispersion, and positive skewness. Economists

Battalio and Kagel received research support from NSF grant GS 32057 and Reynolds thanks the Earhart Foundation for its support. We thank C. G. Miles for making the data for the Cannabis Economy available and R. L. Basmann for first suggesting that we look at income distributions in the experimental economies and for helpful comments on an earlier draft of this paper. Roy Gilbert, Tom Saving, and Bill Brown provided several helpful suggestions and comments. Responsibility for errors remains ours alone.

[*Journal of Political Economy*, 1977, vol. 85, no. 6]
© 1977 by The University of Chicago. All rights reserved.

have not had to look far to find a host of factors which might be responsible for the observed differences in income, although it has proven difficult to discriminate empirically among various hypotheses with available data. Special attention has focused on isolating differences in ability and effort because of a common belief that, if these were the only factors involved in earnings differences, incomes would be more evenly distributed than they currently are. Naturally, many would find it very agreeable if this were the case.

The purpose of this paper is to analyze data on individual labor earnings from two economies in which many of the alleged sources of unequal incomes in national economic systems were greatly reduced and in some cases entirely eliminated. Sufficient experimental control was maintained so that the primary factors reponsible for differences in individual incomes were differences in individual tastes for market income and differences in individual abilities to perform relatively simple job tasks. In other words, each individual could vary his or her level of earnings, subject, of course, to corresponding reductions in leisure time. Under these conditions empirical measures of income dispersion were remarkably similar to comparable figures for the United States and other market economies.

Since this paper describes earnings behavior observed under laboratory-type conditions—an unusual area for economists—we make a few remarks about the relevance of the data. The fact that income dispersion is similar in the experimental and national economies demonstrates that prevailing differences in effort and productivity among relatively homogeneous people (by demographic characteristics) are sufficient to generate differences in earnings as large as those in the U.S. economy, a point which has been seriously questioned (Fair 1971; Blinder 1974). However, this is not the same thing as asserting that the U.S. distribution of income is created solely by individual differences in ability and effort. There are obvious differences between the institutional structure and population characteristics of experimental and national economies which prevent such generalizations. But our results falsify the contention that incomes necessarily will be more evenly distributed in a world where all sources of income differences, except ability and tastes for leisure, are eliminated. In this respect our analysis serves the traditional function of experimentation in the biological and behavioral sciences—to isolate basic relationships under controlled conditions, a technique now receiving a somewhat more respectful hearing in economics (Rivlin 1974; Smith 1976). Moreover, even if it were pronounced impossible to draw reliable inferences about the behavior of a more general population from experiments, we surely would want to know if the empirical regularities reported in national data also occurred under seemingly special conditions.

Section I below describes the experimental economies, Section II describes the variation in individual earnings which occurred under these conditions, Section III presents statistical tests for normality and log-normality in the distribution of earnings, and the last section concludes the discussion.

I. The Experimental Economies

Stigler (1966) provides a representative discussion of factors generating differences in labor incomes, which can be summarized in the following list: (1) luck (personal factors—health, accidents; employer factors—weather, fire, flood, bankruptcy; market factors—general business conditions, finding best job and best pay within a job); (2) direct occupation expenses; (3) costs of occupational training and education; (4) lifetime concavity of earnings with respect to experience; (5) variability or instability of earnings (risk preferences, progressive income tax); (6) differences in cost of living; (7) occupational prestige and social esteem; (8) fringe benefits, like insurance, etc.; (9) monopoly power and restrictions; (10) inability to borrow investment funds for training; (11) imperfect foresight; (12) discrimination and nepotism; (13) relative desire for market income versus leisure; and (14) ability or productive capacity.

The first factor is stochastic; factors 2-8 are the compensatory or equalizing principles of Adam Smith which would exist in full competitive markets and cause wage incomes to differ even among workers of similar ability and taste for income. Factors 9-12 describe divergencies from a perfectly competitive economy which can create additional income differences. The institutional structure of the experimental economies analyzed here permitted factors 1-12 to play, at most, a minor role in accounting for the observed nonuniformity of earnings in these economies. That is, economic conditions were such that differences in ability and preferences for income over leisure were the primary factors responsible for income differences.

Both economies were planned economic environments in the sense that they were organized systems in which individuals lived for continuous, extended time periods and received tokens or points for work performed which were exchangeable for present or future consumption goods.[1] The first was a therapeutic token system established in a female ward for chronic psychotics at the Central Islip State Hospital in New York. The second was an experimental economy (Cannabis Economy) established at the Addiction Research Foundation in Ontario to study the socio-economic effects of marijuana consumption among volunteer subjects.[2]

In the Central Islip economy income was earned primarily in janitorial chores such as sweeping and mopping floors, sorting laundry, and washing dishes, although income could also be earned from self-care activities (straightening up a bed) or activities designed to teach social skills. Each token earning behavior was clearly defined and available at regularly scheduled intervals at a specified token wage rate. Wage rates were structured to approximately clear the labor market, avoiding excess demand or supply in each job activity. Patients did not have fixed job assignments, and the staff allocated janitorial activities among those willing to work on a day-to-day basis. In principle, all patients had the requisite skills to perform any of the jobs, but of course some did higher-quality work than others. Individual patients generally volunteered to perform a given job on a fairly regular basis, and the staff relied upon these patients to perform these jobs. Most patients worked a variety of jobs as well as working the same job more than once a week. However, there was no fixed maximum on the amount of work that an individual was allowed to perform or minimum amount of work required.

---

[1] For a discussion of the use of planned economic environments as laboratories for analysis of economic behavior, see Kagel (1972).

[2] Detailed descriptions of the institutional structures, measurement procedures, and studies of individual consumption and saving behavior in these two economies are available in Battalio et al. (1973) and Kagel et al. (1975).

With respect to consumption opportunities, all patients were supplied with an assured bundle of goods normally provided by a hospital, such as a bed in a public dormitory room, meals, and medical care, independent of token earning activities. Tokens were used to purchase additional goods like cigarettes, coffee, candy, cookies, soda, and milk at the token store, rent a private room or locker, obtain a grounds pass, or purchase clothing and dance tickets. An important feature of consumption behavior in this economy was that the individual distribution of consumption expenditures coincided very closely with the distribution of labor earnings. Examination of saving and consumption behavior in this economy has demonstrated that these consumers conform to standard principles of economic behavior (Battalio et al. 1973; Kagel et al. 1975).

The second economy (Cannabis Economy) consisted of three groups of volunteer subjects, two male groups and one female group, which entered the experiment on different dates. Subjects ranged from 20 to 28 years of age, typically used marijuana (or other derivatives of the cannabis plant) regularly for at least 2 years, were unemployed at the time of volunteering, and had 12 or more years of schooling. The structure of the economic system was the same for all groups except for availability and conditions of marijuana consumption. These differences are ignored here because they did not affect the distribution of output per hour, hours worked, or earnings.[3]

Income-earning activity in the Cannabis Economy consisted of weaving woolen belts on small portable hand looms. Standards for acceptable belts were established and subjects received $2.50 (Canadian currency) for each acceptable belt submitted. The material used in weaving belts was provided free of charge and was always available in the work area. Subjects were free to work anywhere in the ward, during any desired time period, and for as long or as little as they wished. Records of output, earnings, and consumption were maintained by the experimenters, and a record of each subject's activity was made every half-hour.[4]

---

[3] Measures of dispersion in earnings during the predrug phase of each experiment, when subjects did not know which treatment condition they would be assigned to and no cannabis was available, did not differ in any important respect from the measures reported in the text. It appears that the primary effect of cannabis conditions on individual economic behavior lies in the rearrangement of the sequence of daily activities (see Kagel, Battalio, and Miles 1976).

The consumer goods provided free of charge during the 98-day experimental period consisted of assignment, in pairs, to semiprivate rooms with heat, lights, and cleaning services. All other goods had to be purchased from production earnings. A store on the premises offered canteen items, including alcoholic beverages; recreation facilities were available for small rental fees; and meals were available at the hospital cafeteria or snack bar, as well as from food-to-go outlets outside the hospital. A staff member was regularly available to purchase other retail items outside the hospital at market prices, while all unspent earnings could be taken home at the end of the experiment. Consumption and saving behavior in this economy also corresponded to standard principles of economic behavior (Kagel et al. 1975).

II. Differences in Individual Earnings

The data in table 1 provide descriptive measures of the distribution of individual earnings within the two experimental economies. Weekly distributions and individual income totals over 5 weeks are reported for the Central Islip economy.[5] Total earnings for the 3-month experimental

---

[4] An independent check of data accuracy compared the amount of money subjects received upon leaving the economy with what they should have left with on the basis of income and consumption records. Using data from group 3, the differences between these two measures of saving taken as a percentage of individual subject income range from 0 to 6 percent with an average value of 2.4 percent. Independent checks for data accuracy were also maintained in Central Islip by recording total tokens disbursed by weeks and comparing this measure with total earnings. The discrepancy of these two measures over 5 weeks averaged 3.6 percent of the average value of the two measures.

[5] While income is often estimated over a 1-year interval, weekly income distributions are appropriate for the Central Islip economy. Contracting for jobs was daily rather than monthly or annually. Major adjustments to changes in the relative prices of consumption goods were completed within 1 week, as were adjustments to changes in wage rates, a common pattern in token economies. Week-to-week correlations between individual earnings were about 0.9, which are similar to those for annual earnings for national economies. Therefore, it appears that a week is sufficient in a steady-state token economy such as Central Islip to assess the persistent differences among individual earning behaviors. Income is reported for the top 25 earning individuals who were present for all weeks of the study. The rationale for selection was to eliminate patients whose income was primarily earned in self-care activities, i.e., subjects who effectively were nonparticipants in the general labor market. Note also that money wage rates were constant during the reported weeks, but large systematic changes in prices of groups of goods occurred (see Battalio et al. 1973).

TABLE 1

INCOME DISTRIBUTION IN TWO EXPERIMENTAL ECONOMIES

ECONOMY	SEX	N	MEDIAN	MEAN	MIN/MAX	C.V.	VAR. LOGS.	GINI COEFFICIENT	SHARE OF EARNINGS: Lowest* Fifth	SHARE OF EARNINGS: Highest* Fifth	PEARSON'S COEFFICIENT $\frac{3\,(\text{Mean-Median})}{\text{SD}}$	PARETO FUNCTION $\lvert \alpha \rvert$
Central Islip:												
Week 1	F	25	42	47.6	11–157	.80	.51	.406	5.9	45.5	.15	1.13
Week 2	F	25	36	47.5	12–133	.72	.53	.400	5.8	43.1	.34	1.98
Week 3	F	25	44	47.7	13–139	.65	.40	.353	6.6	40.4	.12	1.90
Week 4	F	25	35	42.7	11–170	.87	.44	.402	6.6	47.1	.21	1.07
Week 5	F	25	32	37.9	12–150	.76	.31	.346	8.5	42.5	.20	1.21
5 weeks	F	25	184	223.3	70–689	.68	.34	.344	7.4	41.2	.26	1.38
Cannabis Economy:												
Group 1	M	19	1,857	1,753	720–3,347	.42	.23	.249	9.5	33.1	−.42	3.56
Group 2	M	19	2,385	2,852	1,114–5,397	.45	.22	.264	9.8	35.8	1.06	5.93
Group 3	F	18	1,507	1,600	282–3,747	.60	.56	.358	6.4	42.9	.29	4.12
Groups 1 and 2 ..	M	38	2,105	2,302	720–5,397	.51	.28	.285	8.8	38.5	.50	3.04
Groups 1 ,2, and 3	M-F	56	1,977	2,077	282–5,397	.56	.41	.310	6.6	37.2	.25	3.10

NOTE.—C.V. = coefficient of variation; var. logs. = variance of the logarithm of income

* Lowest and highest 20% rounded to nearest whole number for Cannabis Economy.

period are reported for the three groups in the Cannabis Economy.

Students of income distribution will note some familiar features about the dispersion and skewness shown in table 1. In the Central Islip economy there was considerable dispersion in earnings among individuals in the economy regardless of which index of variability is chosen. The highest earner received over 10 times as much as the lowest earner, the coefficient of variation was about 0.75, the variances of the logarithm of income averaged 0.42, and the Gini coefficient was about 0.375.[6] The absolute value for the slope of the Pareto function was around 1.5, the value which Pareto found long ago. In addition, the Central Islip income distributions all exhibited positive skewness, as indicated by the consistent positive sign for Pearson's coefficient.

The distributions of income from the Cannabis Economy are generally more compact than those for Central Islip. If we consider the combined distribution for all 56 subjects, the coefficient of variation (0.56) and Gini ratio (0.310) were somewhat smaller than those in Central Islip, although the maximum income was over 10 times as large as the minimum income and the variances of the log of income were similar. However, differences in dispersion emerge among the three subgroups comprising this total. Each male group had a maximum income only about five times as large as its minimum income. In addition, coefficients of variation, log variances, and Gini ratios were smaller for the male groups than those reported for the women in table 1. The income distributions in the Cannabis Economy, with the exception of the first group of men, also exhibited positive skewness. The Pareto slopes for the top 20 percent of the Cannabis Economy all exceeded 3.0, indicating that income was distributed more evenly among high income earners in the Cannabis Economy than in Central Islip.[7]

The numerical measures of dispersion reported for the experimental economies fall within the range of comparable measures reported for the United States and other national economies. This is substantiated in table 2, where a sampling of Gini ratios for income distributions in national data is shown. Obviously, no single distribution constitutes *the* correct comparison because representations of income dispersions always depend upon definitions of income recipient unit, income period, and data accuracy, as well as the statistical methods used to summarize and interpret the data. As illustrated in table 2, however, the general tendency in these data is for income to be more evenly distributed the more homogeneous the group of income recipients and the narrower the occupational classifications covered. Gini ratios often exceed 0.40 when distributions include all sources of money income (before personal taxes) for all types of households. However, it is not unusual for Gini ratios of relatively homogeneous groups of workers to be considerably smaller, often falling below 0.30. The Gini indexes for the experimental data fall within these extremes the Cannabis Economy, which had a single source of earnings and a more homogeneous population of workers, had a smaller Gini ratio than the Central Islip economy, which had multiple sources of earnings and a somewhat more heterogeneous population of income earners.

---

[6] The Gini coefficient is defined as the mean difference divided by twice the mean income. In the discontinuous case (individual data) two formulas arise because the mean difference can be calculated with or without repetition. Ours was calculated without repetition, which means that the deviations of each value from itself, which add nothing to the sum of deviations, are not taken, so the number of pairs is $N(N - 1)$ rather than N^2 (see Kendall and Stuart 1969, pp. 46–51). Geometrically the Gini coefficient is the ratio of the area between the Lorenz curve and the diagonal of perfect equality to the total area under the diagonal. The Gini ratio ranges from 0, when incomes are equal (the Lorenz curve coincides with the diagonal), to 1 at the other extreme. Atkinson (1975) offers the following description: Suppose we choose two people at random from the income distribution, and express the difference between their incomes as a proportion of the average income, then this difference turns out to be on average, twice the Gini coefficient: a coefficient of 0.4 means that the expected difference between two people chosen at random is 80 percent of the average income (p. 45).

[7] A Chow test shows that the probability is less than 5 percent that the Pareto value reported from Central Islip for week 2 (the largest Pareto value reported for that economy) is equal to the Pareto value reported for groups 1 and 2 in the Cannabis Economy (the smallest Pareto value reported for that economy).

Comparisons of income dispersion should, however, not rely on any single summary measure, especially since Lorenz curves may cross. But other statistical comparisons suggest the same general conclusion as the Gini index. For instance, the highest fifth of U.S. full-time craftsmen received 32 percent of total earnings in 1970, as did the highest fifth of operatives (Henle 1972). By comparison, the highest fifth of earners in the experimental economies always received *more than* 32 percent of total earnings, indicating as much dispersion of income in the experimental economies as for manual occupations in the national economy. Similarly, Lydall (1968) reported coefficients of variation in labor earnings for all U.S. males of 0.52, 0.36, and 0.37 in 1939, 1949, and 1959, respectively, which are generally smaller than coefficients of variation in the experimental economies.

In summary, data from the experimental economies show that differences in abilities to perform simple manual jobs and differences in preferences for market income among relatively similar individuals (by demographic characteristics) are *sufficient* to generate income dispersions like those observed in national economies. It bears emphasis, however, that this does not imply that these two factors *necessarily* account for all income differences in national data. This issue receives more attention in the conclusion.

III. Functional Form in Distributions

In addition to an interest in dispersion, economists have devoted a great deal of research to determining what the shape of the income distribution

TABLE 2

SELECTED GINI RATIOS FOR INCOME DISTRIBUTIONS IN NATIONAL AND EXPERIMENTAL ECONOMIES

Country or Experiment	Income Concept	Recipient Unit	Gini Ratio
1. United States	IRS adjusted gross income	Individual tax returns	.46
2. Central Islip	Labor earnings (no personal tax)	Female workers	.41 (max)
3. United States	Pretax money income	Families and unattached individuals	.40
4. United States	Pretax money income	Female-headed families	.40
5. United States	Income after taxes and transfers	Families and unattached individuals	.36
6. United States	Pretax money income	Families	.35
7. Central Islip	Labor earnings (no personal tax)	Female workers	.34 (min)
8. Cannabis	Pretax labor earnings	All workers	.31
9. United States	Consumption expenditures	Families and unattached individuals	.31
10. India	Consumption expenditures	Individuals	.31
11. United States	Pretax labor earnings	Household heads, age 25–34	.29
12. Cannabis	Pretax labor earnings	Male workers, age 20–28	.29
13. United States	Pretax wages and salaries	Male workers, full time, all industry	.28
14. United States	Pretax wages and salaries	Male workers, full time, manufacturing	.26
15 United Kingdom	Pretax wages and salaries	Male workers, full time employees	.21

SOURCES.—For national estimates, line 1. Gastwirth 1972, line 3 U.S Bureau of the Census 1974*a*, line 4: Danziger and Plotnik 1975; lines 5 and 11: Taussig 1973; line 6: U.S. Bureau of the Census 1974*b*, also see Paglin 1975 and Taussig 1976, line 9 Reynolds and Smolensky 1977, line 10 Paukert 1973; lines 13 and 14 Henle 1972; line 15: Hill 1959.

is like, especially since this could provide the basic datum which a theory of income distribution must explain. Three functional forms for the distribution of income have been especially prominent: the Pareto distribution, the normal distribution, and the lognormal distribution (Lydall 1968, p. 13). The Pareto distribution provides a reasonably adequate characterization of earnings for the top 20 percent of the income distribution in the experimental economies, with simple correlation coefficients ranging from .87 to .98 for the Pareto functions reported in table 1. This is typically the case for national data as well, although the Pareto distribution fits overall distributions poorly.

The hypothesis of normality in the distribution of earnings has been a candidate to fit the overall distribution because of the possibility that human earning capacities are normally distributed, just as some other variates in nature are. The second major candidate has been the lognormal distribution, which has a long history based upon statistical models like the law of the proportionate effect or the multiplicative interaction among abilities relevant to producing earnings. The *W*-statistic (Shapiro and Wilk 1965) was used to test the hypothesis that earnings were distributed normally (or lognormally) in the experimental economies, as Monte Carlo studies have shown it to be quite powerful for small sample sizes relative to other common tests (Shapiro, Wilk, and Chen 1968).

The hypothesis of lognormality in the distribution of earnings was accepted at the 1 percent confidence level for each week and for incomes totaled over 5 weeks in the Central Islip economy. By contrast, the hypothesis of normality was rejected at this confidence level.

The Cannabis data did not conform quite so clearly to only one functional form. Lognormality was accepted for the distribution of income in all three experimental groups, but normality was also accepted at the 1 percent level. To resolve this ambiguity we reorganized the data to see if one functional form would dominate. When the earnings for all men were pooled (experiments 1 and 2), lognormality was accepted and normality rejected. Similarly, in pooling the data for all subjects, lognormality was accepted and normality rejected at the 1 percent level.

The results of our tests for normality and lognormality are reasonably similar to those using data from national economies (Lydall 1968). The hypothesis of normality is usually in poor agreement with national data but provides a somewhat better characterization in more narrowly defined manual occupations. Similarly, normality did not fit as badly in the Cannabis groups as in Central Islip, which had a more heterogeneous job structure. As with most national data, the hypothesis of lognormality was in better agreement with the data from both experimental economies than the hypothesis of normality.

We note one more striking similarity between the Cannabis data and national data, namely, the difference between male and female earnings. Lydall (1968, p. 242), among many others, has puzzled over this ancient difference between the sexes: "For full-period workers in the developed countries, the median earnings of women seem to be mostly between a half and two-thirds of the earnings of men; and the question which arises is whether this ratio is an accurate reflection of the differences of average effective abilities of men and women, or whether it is partly institutionally or sociologically determined." Remarkably enough, female mean earnings in the Cannabis experiment were 69.5 percent of the mean earnings for all males, a statistically significant difference at the 1 percent level. Many explanations might account for the observed difference—differences between the sexes in physical stamina or desire for leisure—but, given the structure of the Cannabis Economy, it cannot be accounted for by sex bias or segregation of the sexes into different jobs.

IV. Conclusions

We have analyzed data on the distribution of incomes in two experimental economies in which many of the factors commonly cited as responsible for generating income differences were absent. Nonetheless, measured income dispersion and functional form of the distributions in these simple economies is not very different from those in the United States and other market economies.

Since the bulk of empirical work on the distribution of incomes generally does not derive from analytical models in which the income distribution is either an important dependent or independent variable, much of the interest appears to be motivated ideologically. Some might be tempted to use our results on behalf of various opinions about the fairness of the existing distribution of income. We make only two points. The first is that this study provides evidence for the proposition that measures of dispersion in income and earnings reveal virtually nothing about the socioeconomic structure in which the observed distribution was generated. We considered two institutional structures and populations that were quite different from national economic systems and from each other, yet the empirical measures of income dispersion were not dramatically different. This draws attention to the fact, sometimes overlooked, that particular individuals occupying a given income rank may be quite different under alternative economic structures although the overall distribution remains unchanged. Thus, in an economic system where factors such as 9 and 12 (monopoly power, nepotism, discrimination) are important factors generating a particular distribution of income, removal of these factors need not change the empirically measured Gini coefficient or coefficient of variation but certainly can change the income position of individuals.

The second point is that our study provides evidence for the importance of differences in individual tastes for market income and abilities as major determinants of the income distribution. For instance, our evidence certainly throws doubt upon the empirical relevance of the theoretical model which led Fair (1971, p. 575) to conclude that "it is thus apparent that one has to make rather extreme assumptions about the variation of the productivity and industriousness of people before the optimal Gini coefficient approaches 0.35." But even if preferences and productive abilities play a dominant role in creating the income dispersion observed in the U.S. economy, it does not follow that the distribution of income is immutable. To the extent that tastes and abilities are formed by environmental experience, changes in these variables potentially can modify the equilibrium distribution of income.

References

Atkinson, A. B. *The Economics of Inequality*. London: Oxford Univ. Press, 1975.

Battalio, R. C.; Kagel, J. H.; Winkler, R. C.; Fisher, E. B., Jr.; Basmann, R. L.; and Krasner, L. "A Test of Consumer Demand Theory Using Observations of Individual Consumer Purchases." *Western Econ. J.* 11 (December 1973): 411-28.

Blinder, A. S. *Toward an Economic Theory of Income Distribution*. Cambridge, Mass.: M.I.T. Press, 1974.

Danziger, S., and Plotnik, R. "Demographic Change, Government Transfers, and the Distribution of Income." *Monthly Labor Rev.* 100, no. 4 (April 1977): 7-11.

Fair, R. C. "The Optimal Distribution of Income." *Q.J.E.* 85 (November 1971): 551-79.

Gastwirth, J. L. "The Estimation of the Lorenz Curve and Gini Index." *Rev. Econ. and Statis.* 54 (August 1972): 306-16.

Henle, P. "Exploring the Distribution of Earned Income." *Monthly Labor Rev.* 95, no. 12 (December 1972): 16-27.

Hill, T. P. "An Analysis of the Distribution of Wages and Salaries in Great Britain." *Econometrica* 27 (July 1959): 355-81.

Kagel, J. H. "Token Economies and Experimental Economics." *J.P.E.* 80, no. 4 (July/August 1972): 779-85.

Kagel, J. H.; Battalio, R. C.; and Miles, C. G. "The Effects of Cannabis Consumption on Production, Hours Worked, and Efficiency: An Experimental Investigation." Mimeographed. College Station: Texas A&M Univ., 1976.

Kagel, J. H.; Battalio, R. C.; Winkler, R. C.; Fisher, E. B., Jr.; Basmann, R. L.; and Krasner, L. "Income Consumption and Saving Behavior in Controlled Environments: Further Economic Analysis." In *Experimentation in Controlled Environment*, edited by C. G. Miles. Toronto: Addiction Res. Found., 1975.

Kendall, M. G., and Stuart, A. *The Advanced Theory of Statistics*. New York: Hafner, 1969.

Lydall, H. *The Structure of Earnings*. London: Oxford Univ. Press, 1968.

Okun, A. M. *Equality and Efficiency: The Big Trade-off*. Washington: Brookings Inst., 1975.

Paglin, M. "The Measurement and Trend of Inequality: A Basic Revision." *A.E.R.* 65 (September 1975): 598-609.

Paukert, F. "Income Distribution at Different Levels of Development: A Survey of Evidence." *Internat. Labour Rev.* 108 (August/September 1973): 97-125.

Reynolds, M. O., and Smolensky, E. *Public Expenditures, Taxes, and the U.S. Distribution of Income.* New York: Academic Press, 1977.

Rivlin, A. M. "How Can Experiments Be More Useful?" *A.E.R., Papers and Proc.* 64 (May 1974): 346-54.

Shapiro, S. S., and Wilk, M. B. "An Analysis of Variance Test for Normality (Complete Samples)." *Biometrika* 52 (December 1965): 591-611.

Shapiro, S. S.; Wilk, M. B.; and Chen, H. J. "A Comparative Study of Various Tests for Normality." *J. American Statis. Assoc.* 63 (December 1968): 1343-72.

Smith, V. L. "Experimental Economics: Induced Value Theory." *A.E.R., Papers and Proc.* 66 (May 1976): 274-79.

Stigler, G. J. *The Theory of Price.* 3d ed. New York: Macmillan, 1966.

Taussig, M. K. *Alternative Measures of the Distribution of Economic Welfare.* Princeton, N.J.: Princeton Univ. Press, 1973.

———. "Trends in Inequality of Well-Offness in the United States since World War II." Mimeographed. New Brunswick, N.J.: Rutgers- The State Univ., 1976.

U.S. Bureau of the Census. *Current Population Reports.* Series P-60, no. 89. Washington: Government Printing Office, 1974. (*a*)

———. *Current Population Reports.* Series P-60, no. 90. Washington: Government Printing Office, 1974. (*b*)

Health, Investment in Health, and Occupational Choice

M. L. Cropper

University of California, Riverside

This paper presents two models of investment in health which explicitly recognize the random nature of illness and death. The first model examines life-cycle behavior of investment and health capital when the motive for investing in health is to decrease the probability of illness. In the second model the individual invests in health through his choice of occupation. This determines the extent of his exposure to a pollutant, such as asbestos, which increases the probability of death. The model examines how exposure to pollution should vary with age and predicts how workers should respond to information about occupational dangers.

I. Introduction

In recent years it has been recognized that expenditures on medical services, nutrition, and exercise can be viewed as investments in health capital and analyzed using the tools of capital theory. This approach has enabled economists to derive propositions about the pattern of medical expenditures over an individual's lifetime and to describe the behavior of health capital over the life cycle. In addition, it has provided a means of explaining observed cross-sectional differences in medical expenditures. Grossman (1972), for example, has developed a model in which health is viewed as a capital stock which yields an output of "healthy days." Individuals may invest in health by combining time (e.g., for doctor's visits) with purchased inputs (e.g., medical services). The incentive for investing in health is that by increasing the health stock the individual increases the amount of time available for earning income or for producing consumption goods.

The author would like to thank Michael Grossman, Tracy Lewis, and an editor of this *Journal* for helpful comments on an earlier draft of this paper while retaining full responsibility for any errors.

[*Journal of Political Economy*, 1977, vol. 85, no. 6]
© 1977 by The University of Chicago. All rights reserved.

In the present paper, investment in health is approached from a slightly different viewpoint. Specifically, I assume that individuals invest in health not to receive a monetary return but to avoid the disutility associated with being ill.[1] Since individuals are usually unable to enjoy life fully when they are ill, illness will be treated as a temporary interruption in the individual's utility stream. Death, which is also affected by the stock of health capital, can be regarded as a permanent interruption of the utility stream.

A second point of departure is that the relation between health capital and illness is treated here as random rather than as deterministic. Whether one is well or ill depends, after all, on random events—changes in climate, exposure to viruses and germs—as well as upon the size of one's health stock. Formally, I shall assume that illness occurs whenever the health stock, H, falls below some critical level $\bar{H}$, where $\bar{H}$ is a random variable. The individual thus increases the probability of being well by maintaining a high stock of health capital but cannot guarantee that illness will not occur. An equivalent way of expressing this is to say that at each instant of the individual's life one of two states—"ill" or "not ill"—must occur. The probability of either state obtaining depends on the stock of health capital and can be determined indirectly by investing in health.

Since individuals are not fully able to enjoy life when they are ill, illness is assumed temporarily to interrupt the individual's utility stream; that is, if the state "ill" occurs the individual receives the same utility he would receive if he consumed nothing. Thus psychically, if not literally, illness is equivalent to zero consumption.

Aside from possible theoretical appeal, this treatment of illness enables one to distinguish between two very different types of medical expenditure: preventative medical expenditure and expenditure in the event of illness. This distinction is an important one since the two types of investment may behave quite differently over the life cycle. The present model focuses on preventative expenditure on health, which is nonstochastic, and shows that this form of investment may decrease as the individual ages and the period over which returns are received decreases. Expenditure in the event of illness, on the other hand, is a random variable which is positively related to the occurrence of illness. Although the model does not deal explicitly with this type of investment, it is shown that the fraction of time spent ill increases as the individual ages, hence so should expenditure necessitated by illness.

A second advantage of our approach to investment in health is that it provides a model which, with suitable modification, can be used to analyze a more subtle form of investment in health capital. Casual observation suggests that there is often a trade-off between high wages and job safety, especially in unskilled jobs. Hence, in order to earn a higher income an individual must often work in occupations which are damaging to his health and may even result in his death. This implies that the choice of an occupation is also a form of investment in health.

---

[1] Grossman emphasizes that good health should be regarded as a consumption as well as an investment good; however, in the principal case which he analyzes, health services do not enter the utility function directly.

In recent years attention has been focused on a particular class of jobs which are injurious to the health, namely, jobs involving industrial pollutants.[2] The more well-publicized of these pollutants include asbestos, kepone, and nuclear radiation; however, the problem is considerably larger, for the U.S. Food and Drug Administration lists some 25,000 substances as potentially hazardous to the health. Like asbestos, most industrial pollutants are absorbed gradually, in proportion to the number of hours worked, and accumulate inside the worker's body. The interesting feature of the process is that the effects of these pollutants may not be felt for many years; however, once enough of the pollutant has accumulated, death will result. Unfortunately, what constitutes a lethal dose of many pollutants must be considered uncertain. The problem facing the worker is therefore one of trading higher wages for an increase in the probability of death.

The existence of jobs involving industrial pollutants raises several interesting questions. First, if an individual is faced with a choice between a job such as coal mining which is known to be dangerous and a safer but lower-paying occupation, how should he optimally behave? That is, by how much should he increase the probability of death in order to earn a higher income, and how should employment in the risky occupation vary over the life cycle? Second, if an individual has been working for years in an occupation which is suddenly discovered to be dangerous to the health, should he continue working, given that he has already accumulated a large stock of the pollutant, or should he change his occupation and allow the pollutant to decay?

The answers to these questions, which are developed in Section III, are somewhat surprising. It might seem, for example, that if an individual chooses to work at all in a risky occupation, he should do so when he is young, switching to a healthier occupation as he grows older so that his pollution stock may decrease. Under plausible conditions, however, it may be shown that the optimal pollution stock never decreases and that some exposure to the pollutant is always called for. Similarly, when an individual who has inadvertently accumulated a large stock of pollutant suddenly discovers that the substance is dangerous to his health, optimal behavior may consist of further increasing the stock of the pollutant rather than switching to a less-risky job.

---

[2] A second type of job-related health risk is the risk of industrial accidents. Unlike pollutants, whose effects persist long after employment has ended, these risks are of a temporary nature only and may be thought of as increasing the probability of death while the individual is at work. Since the problem of industrial accidents has been dealt with by Thaler and Rosen (1975), I shall not deal with that problem here.

These conclusions are based on a model, set forth in Section III, which is formally analogous to the model of conventional investment in health presented in Section II. Results are summarized in Section IV.

II. A Model of Investment in Health Capital

A. The Model

Let us assume that at each instant an individual receives an income of $\bar{Y}$ which he divides between expenditure on a consumption good, C, and investment in health capital. This assumption ignores the possibility that the individual invests in human or nonhuman capital; however, it permits us to focus on investment in health and reduces the model to a single state variable. If investment in health has a constant marginal cost of c and if the consumption good has a price of v per unit, then the instantaneous budget constraint may be written

$$\bar{Y} = cI + vC, \tag{1}$$

or equivalently,

$$C = Y - pI, \tag{2}$$

where $Y \equiv \bar{Y}/v$ and $p \equiv c/v$.[3] Expenditure on consumption increases the individual's utility, provided that he is well, while investment in health increases the stock of health capital according to the relation

$$\dot{H} = I - \delta H. \tag{3}$$

Equation (3) reflects the fact that the health stock decays over time as the individual ages. Initially it is assumed that the rate of decay is constant throughout the individual's lifetime; however, the possibility that δ increases over time is considered below.

The size of the health stock is important since it determines at any instant whether the individual is well or ill. Specifically, the event "illness" occurs whenever the health stock H_t falls below a critical sickness level $\bar{H}_t$, which is assumed to be random. If $\bar{H}_t$ is viewed as the level of germs and viruses to which the individual is exposed, and if H_t is interpreted as resistance to disease, then this assumption characterizes fairly accurately minor illnesses such as colds, viruses, and influenza. In keeping with this interpretation of illness it is assumed that sickness lasts only momentarily and does not reduce the stock of health capital.[4]

---

[3] Another way of expressing this is to say that at each instant the individual has a fixed amount of labor which he can devote to earning income, producing consumption goods, or producing investment in health. Consumption and investment in health are produced, as in Grossman (1972), by combining labor with purchased inputs. As long as the production functions for C and I are homogeneous of degree one, this production-oriented framework is equivalent to the assumptions made above.

The random nature of $\tilde{H}_t$ is captured in the assumption that there exists a density function $f(\tilde{H})$ over the interval $[0, \infty)$ from which a new $\tilde{H}$ is drawn at each t. Successive drawings are assumed independent so that $\tilde{H}_t$ is independently and identically distributed for all t. In view of these assumptions, the probability of the state "not ill" may be written

$$F(H_t) = \int_0^{H_t} f(\tilde{H})\, d\tilde{H},$$

while the probability of being ill is given by $1 - F(H_t)$.

An important feature of the model is the way in which illness affects the individual's utility. If at any t the state "well" occurs, then the individual receives utility from consumption according to the function $U(C)$,

$$U'(C) > 0,\ U''(C) < 0,\ U'(0) = \infty,\ U(0) = 0. \tag{4}$$

If the state "illness" occurs, utility is given by the function $V(C) \equiv 0$, all C. Although the individual may literally consume some of the good C when he is ill, it is assumed that the disutility of illness renders this psychically equivalent to zero consumption.[5]

Since illness is a random event, instantaneous expected utility can be written, by substitution from (2), as

$$EU_t = F(H)U(Y - pI). \tag{5}$$

Given an initial stock of health H_0, the problem facing the individual is to choose a path I_t and a value H_T to maximize

$$\int_0^T e^{-\rho t} F(H)U(Y - pI)\, dt \tag{6}$$

---

[4] Since the model applies only to minor illnesses which do not affect the stock of health capital, it is reasonable that the individual plans preventative health expenditure at $t = 0$ for the remainder of his life. If one wanted to model major illness, this could be accomplished by having illness (a random event) reduce the size of the health stock. In this case H_t would be a random variable, and the individual would determine his investment in health at each t, based on the most recently observed value of H_t.

[5] The assumption that $V(C) \equiv 0$ in the event of illness does not imply that the individual will plan to consume nothing when he is ill. For further discussion about this point the reader is referred to the beginning of Section IIB. One might criticize my treatment of illness on the grounds that it does not distinguish among illnesses according to their severity. This can be corrected, however, by assuming that the individual receives utility from services ϕH_t which are proportional to the stock of health. If the individual is well, utility is given by the strictly concave and increasing function $U(C, \phi H)$, while utility in the event of illness is assumed to be $U(0, \phi H)$. (U[Death] = 0, as in the text.) If one also assumes that $U(C, \phi H)$ is separable, then the qualitative results obtained in Section IIB are unaltered by including H in the utility function.

subject to (3).[6] It is also reasonable to assume that $I \geq 0$ since health capital can neither be sold nor consumed.

In equation (6), T, the time of the individual's death, is assumed known with certainty and outside of his control. This rather restrictive assumption is relaxed below; however, the case in which T is fixed provides an interesting basis for comparing investment in health with investment in human capital.

B. *Life-Cycle Behavior of Investment and Health Capital*

In what follows I shall describe the paths of investment and health capital by examining the necessary conditions corresponding to equation (6). Before doing so, however, it should be established that the path of investment which maximizes (6) is consistent with the amount of investment which the individual would choose at each t based on the most recent information available to him. I shall argue that the path of planned investment is consistent with the actual path of investment if the distribution $F(\bar{H})$ is regarded as an objective distribution and if purchases of consumption goods at instant t must be made ε minutes before the individual knows whether or not he is ill.[7]

Suppose, for example, that the individual revises his investment decision at time τ, given knowledge of the path $\bar{H}_t$, $0 \leq t < \tau$. If $F(\bar{H})$ is regarded as the true distribution from which a value of $\bar{H}$ is drawn at each t, then the individual has no reason to alter $F(\bar{H})$ at time τ. If it is also assumed that the individual must order the goods which he will consume at τ an arbitrarily small amount of time before $\bar{H}_\tau$ is revealed, then the individual at τ will solve the problem

$$\max_{I_t} \int_\tau^T e^{-\rho(t-\tau)} F(H) U(Y - pI)\, dt, \tag{7}$$

given H_τ, and the values of C_τ and I_τ chosen at τ will not differ from the values of C_τ and I_τ chosen at time 0.

The paths of investment and health capital may now be examined by writing the Hamiltonian function and necessary conditions corresponding

---

[6] If the instantaneous utility function $U(C)$ is bounded from below, then by Tonelli's theorem it can be shown that equation (6) is equivalent to the objective function suggested by the expected utility hypothesis,

$$E\left[\int_0^T e^{-\rho t} V(t)\, dt\right],$$

where

$$V(t) = \begin{cases} U(C_t) & \text{if } H_t \geq \bar{H}_t \\ 0 & \text{if } H_t < \bar{H}_t \end{cases}.$$

[7] There are many cases in which this assumption seems to fit reality. Food and clothing, for example, are goods which are usually purchased prior to the instant of consumption and cannot easily be resold.

to equation (6). Omitting the discount factor, the current-value Hamiltonian becomes

$$\mathcal{L} = F(H)U(Y - pI) + \mu(I - \delta H) + rI, \tag{8}$$

where μ is the multiplier attached to (3) and r corresponds to the constraint $I \geq 0$. For any H and μ, I must maximize $\mathcal{L}$, hence (9) must hold along an optimal path. Equations (10)–(13) must also be satisfied.

$$\partial\mathcal{L}/\partial I = -F(H)pU'(Y - pI) + \mu + r = 0, \tag{9}$$

$$\dot{\mu} = (\rho + \delta)\mu - F'(H)U(Y - pI), \tag{10}$$

$$\dot{H} = I - \delta H, \tag{11}$$

$$e^{-\rho T}\mu_T H_T = 0, \tag{12}$$

$$r \geq 0, rI = 0. \tag{13}$$

Since equations (9)–(13) play an important role in subsequent analysis, it is well to examine them before proceeding. In these equations, μ, the shadow price of health, represents the increase in future expected utility of a unit increase in the stock of health capital. Equation (9), which determines I as a function of μ and H, states that the individual will increase current consumption at the expense of investment in health capital until the expected marginal utility of a unit of current consumption just equals μ, the future utility foregone by not investing in health.

Equation (10), which may also be written

$$F'(H)U(Y - pI) = \left(\rho + \delta - \frac{\dot{\mu}}{\mu}\right)\mu, \tag{14}$$

is a condition which must be satisfied by the optimal stock of health capital. In interpreting (14) recall that for the individual, investment in health is analogous to investment in a marketable capital good such as a consumer durable. What the equation says is that $F'(H)U(Y - pI)$, the increase in utility obtained by purchasing an additional unit of H for an instant, must equal $[\rho + \delta - (\dot{\mu}/\mu)]\mu$, the cost of an instantaneous increase in H. This cost is comprised of three parts—the psychic return ρ foregone by "spending" μ on H rather than on current consumption; the depreciation cost $\delta\mu$, reflecting the instantaneous deterioration in H; and $\dot{\mu}$, the "capital gain" which accrues to the holder of H. (The latter must be subtracted from $[\rho + \delta]\mu$ to arrive at the cost of capital.) Equation (10) is thus analogous to the familiar result from capital theory that the marginal product of capital must equal the supply price of capital—interest plus depreciation minus capital gains. The difference here is that because the stock of health is not marketable, all returns must be measured in utility terms.

Equations (9)–(11) must be satisfied by I, H, and μ at all points along an optimal path. Equation (12), on the other hand, places restrictions on these variables at the terminal date T. Equation (12) indicates that at the time of his death an individual will optimally have used up his stock of health capital or will be in a position where further increases in the stock of health cannot increase utility ($\mu_T = 0$). It is clear, however, from (11) and the fact that I must be nonnegative, that the individual will never be able to deplete his stock of health capital in finite time. It therefore follows that the shadow price of health must be zero at the time of death. Since μ is continuous, this implies that $\mu < F(H)pU'(Y)$ for some interval prior to T, and hence that $I = 0$ prior to the time of death.

Thus as long as the time of death is known with certainty, it must be preceded by a period in which no expenditures on preventative health care are made. This observation also implies that the individual's health stock will be decreasing steadily toward the end of his life and that illness is likely to occur more frequently prior to death.

We shall now examine how investment in health and health capital change over the individual's lifetime. Equations (9)–(13) do not exclude the possibility that the individual never invests in health; however, the case of $I_t > 0$ (except at the time of death) is clearly more interesting. When $I_t > 0$ the rate of change in investment, $\dot{I}$, may be derived by differentiating equation (9) with respect to time and then substituting for $\dot{\mu}$ and $\dot{H}$ from equations (10) and (11).

$$\dot{I} = \frac{-U'}{pU''}\left[\delta + \rho - \frac{F'(H)}{F(H)}\left(\frac{U}{pU'} + I - \delta H\right)\right]. \tag{15}$$

Equations (11) and (15), when solved simultaneously, yield a family of I-H paths, one for each value of T. The properties of these paths are examined in figure 1 by plotting the stationary loci $\dot{I} = 0$ and $\dot{H} = 0$.

If $\dot{H}$ is set equal to zero, the result is a straight line with slope $1/\delta$. For values of I to the left of the line, $\dot{H}$ is decreasing while $\dot{H} > 0$ to the right. The shape of $\dot{I} = 0$ is, however, more complicated. For $I > 0$ the locus $\dot{I} = 0$ may conveniently be written

$$\begin{aligned}\frac{U(Y - pI)}{pU'(Y - pI)} + I &= \frac{\delta + \rho}{\xi(H)} + \delta H \\ &\equiv J(H),\end{aligned} \tag{16}$$

where $\xi(H) \equiv F'(H)/F(H)$ is the conditional probability density of illness. Loosely interpreted, this is the probability that the individual falls ill when his health stock is H_t, given that he is well when his health capital exceeds H_t. We shall assume, as seems reasonable, that the conditional probability of illness is a decreasing function of H and that

$$\infty \geq \lim_{H \to 0} \xi(H) > 0, \quad \text{and} \quad \lim_{H \to \infty} \xi(H) = 0. \tag{17}$$

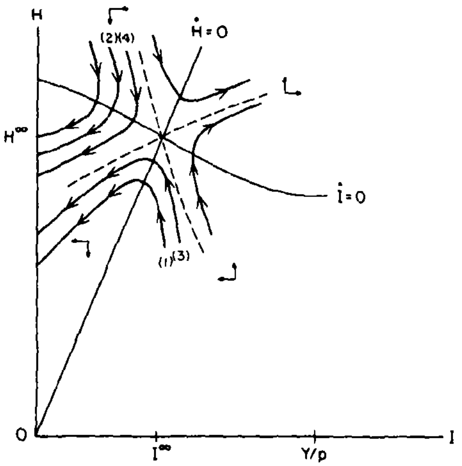


FIG. 1.– Paths of investment and health capital when death is exogenous

In view of this assumption, the right-hand side of (16) is a strictly increasing function of H. As the reader may verify, the left-hand side of the equation is a decreasing function of I for $0 \leq I \leq Y/p$ and an increasing function of I for $I > Y/p$.[8] The locus $\dot{I} = 0$ therefore begins at the point $\{0, J^{-1}[U(Y)/pU'(Y)]\}$, slopes downward to the point $[Y/p, J^{-1}(Y/p)]$, and increases thereafter. It can also be shown that the loci $\dot{H} = 0$ and $\dot{I} = 0$ must intersect only once, and that this intersection must occur on the downward-sloping portion of $\dot{I} = 0$.[9]

We are now in a position to describe the optimal paths of investment and health capital. In figure 1 the heavy solid lines represent solutions to the system (11)–(15) for various values of T. Paths such as (3) and (4), which more closely approach the stationary values I^{∞} and H^{∞} than do paths (1) and (2), correspond to larger values of T. Since investment in health must be zero at the time of death, any path which does not terminate on the H-axis cannot be optimal. This implies that an optimal path will resemble (1) or (3) if H is initially small ($H_0 < H^{\infty}$) and (2) or (4) if H_0 is large ($H_0 > H^{\infty}$).

---

[8] Specifically, the derivative with respect to I of the left-hand side of (16) is $UU''/p(U')^2$, which is negative for $U > 0$ and positive for $U < 0$.

[9] Whenever the curves $\dot{I} = 0$ and $\dot{H} = 0$ intersect, it must be the case that $I = \delta H$ and $\delta + \rho = \xi(H)[(U/pU') + I - \delta H]$. Substituting the first equation into the second and rearranging terms, $\xi(H) = (\delta + \rho)p[U'(Y - p\delta H)]/[U(Y - p\delta H)]$, which may be solved for H. The left-hand side of the equation is, by assumption, monotone decreasing in H, while the right-hand side is strictly increasing in H. It thus follows that the stationary loci can intersect only once. Since the right-hand side of the equation approaches infinity as H approaches $Y(\delta p)^{-1}$, this implies that $H^{\infty} < Y(\delta p)^{-1}$ and hence that $I^{\infty} < Y/p$. Hence the phase diagram of the system must appear as is figure 1.

"Unhealthy" individuals (those for whom $H_0 < H^\infty$) will therefore build up their stock of health capital during their early and middle years and let it decay during old age. This will be achieved by investing heavily in health in the early years, presumably via nutrition and exercise, with I decreasing steadily throughout the individual's lifetime and reaching zero sometime before death.

What is perhaps a more typical case occurs if the individual is in good health at the beginning of his lifetime ($H_0 > H^\infty$). The optimal plan in this instance is to allow the stock of health capital to decrease steadily over time. The corresponding path of investment will not, however, be monotone decreasing. For T sufficiently large, the optimal strategy is to increase the amount invested in health to a peak in the early years of life and let it decline steadily thereafter. This initial buildup of I is necessary to insure that the health stock does not decline too fast for long-lived individuals.

The results of figure 1 suggest an interesting parallel between investment in health capital and investment in human capital. In the human-capital literature (see Blinder and Weiss 1976), the time devoted to acquiring productive skills usually declines over the life cycle as the period over which returns are received decreases. This phenomenon also occurs above. Investment in health decreases as death approaches and the marginal benefits from increasing the stock of health capital (μ) decline. Regarding the behavior of human capital, the individual usually builds up his stock of productive skills early in life, with the stock reaching a peak in the middle years and decreasing thereafter. This is also true of health capital if the initial endowment of health is low ($H_0 < H^\infty$). However, since the decision to invest in health is presumably made during adolescence or early adulthood, it is very likely that the stock of health capital at that time will be high and will decline steadily throughout the individual's lifetime.

It should also be noted that the paths of investment in figure 1 contrast sharply with the results obtained by Grossman (1972) regarding life-cycle behavior of investment in health. A point emphasized by Grossman is that investment in health will be positively correlated with age as long as the elasticity of the marginal efficiency of health capital is less than one. This conclusion is reasonable if investment in health is interpreted to mean expenditure in the event of illness, since individuals are more likely to be ill when they are old than when they are young. It is less likely, however, that *preventative* medical expenditures will increase over time, since investments made late in life yield returns over a shorter period than investments made early in life. In the context of Grossman's model a portion of investment in health may be interpreted as recuperative expenditure, and hence it is reasonable that investment in his model should increase over time. In my model, on the other hand, investment in health must be interpreted solely as preventative health care, since expenditure in the event of illness is a random event. Hence it is reasonable that I decreases over the life cycle.

C. Investment in Health When Death Is Endogenous

It should be remembered that all of the preceding results rest on the somewhat restrictive assumption that the stock of health capital in no way influences the time of death. A more reasonable assumption is that death occurs when the health stock falls below some level $\hat{H}$, where $\hat{H} \leq \tilde{H}$. This definition of death requires, however, that the assumptions regarding $\tilde{H}$ be altered slightly. It is now assumed that $\tilde{H}$ is independently and identically distributed for all t over the interval $[\hat{H}, \infty)$.[10] Death, which occurs when H_t falls below $\hat{H}$, is therefore equivalent to being sick 100 percent of the time. This assumption may conflict with the observation that some persons prefer being invalids to dying; however, it is hard to define death in a more "realistic" way which is also analytically convenient.

If this definition of death were incorporated into the model above, T would be determined by the condition $H_T = \hat{H}$, and the individual would clearly choose to live forever. The choice of a finite lifetime can, however, be explained by allowing the rate of depreciation of the human capital stock to increase as the individual ages, a device first employed by Grossman (1972). If the health stock decays fast enough, desires for present consumption, together with the individual's budget constraint, will cause him to live a finite life.

Unfortunately, if δ is a function of time, the model of Sections A and B is no longer autonomous and the analysis becomes complicated. It is nevertheless possible to say something about the paths of investment and health capital for the case in which the individual is risk neutral. Let us therefore assume that the individual solves the control problem

$$\max_{I_t, T} \int_0^T e^{-\rho t} F(H)(Y - pI)\, dt \tag{18}$$

subject to the constraints $\dot{H} = I - \delta H$, $0 \leq I \leq Y/p$, and $H_T = \hat{H}$.

Since the Hamiltonian function

$$\mathscr{L} = F(H)(Y - pI) + \psi(I - \delta H) \tag{19}$$

---

[10] The expression $F(H)$ is now redefined as

$$\int_{\hat{H}}^{H_t} f(\tilde{H})\, d\tilde{H}.$$

is now linear in I, the individual will set

$$\begin{aligned} I &= 0 \quad \text{if} \quad \psi - pF(H) < 0, \\ I &= \frac{Y}{P} \quad \text{if} \quad \psi - pF(H) > 0, \end{aligned} \tag{20}$$

where ψ is the value of a unit increase in H and $\psi - pF(H)$ may be interpreted as the net yield, measured in utility terms, of an additional unit of investment. In the intermediate case in which $\psi - pF(H)$ is zero, any value of I in the interval $[0, Y/p]$ is possible. The optimal I is determined in this case by differentiating $\psi - pF(H)$ with respect to time,

$$\frac{d}{dt}[\psi - pF(H)] = \dot{\psi} - pF'(H)\dot{H} = 0, \tag{21}$$

and substituting for $\dot{H}$ and $\dot{\psi}$ from equations (3) and (22), both of which must hold along an optimal path. The resulting equation, (23), may be solved for

$$\dot{\psi} = \psi(\delta + \rho) - F'(H)(Y - pI), \tag{22}$$

$$\frac{p(\delta + \rho)}{Y - p\delta H} = \xi(H), \tag{23}$$

and the optimal health stock, H_t^*, and investment, I_t^*, may be found in turn from the state equation $I = \delta H + \dot{H}$.[11]

The important question, of course, is how H_t^* and I_t^* vary over the life cycle. As in Section IIB the analysis below is confined to the case in which $I_t > 0$, except at the time of death. To describe the behavior of health capital over time, equation (23) is graphed in figure 2. The left-hand side of the equation is monotone increasing in H and discontinuous at $H = Y(p\delta)^{-1}$, while the right-hand side is a decreasing function of H which either intersects or is asymptotic to the line through $\hat{H}$. Since initially δ will be small, it is likely, unless the cost of investment is very high, that $Y(p\delta)^{-1}$ will exceed $\hat{H}$, thus guaranteeing a unique solution for H^*. As the individual ages and his health capital decays at a faster rate, $Y(\delta p)^{-1}$ grows smaller and $p(\delta + \rho)(Y - p\delta H)^{-1}$ shifts graphically upward. Thus, as expected, the optimal stock of health capital decreases monotonically, with the precise rate of decrease given by

---

[11] In determining the paths I_t^* and H_t^*, a problem arises if the solution to (23) is not unique or if the value of I_t^* implied by (23) violates the constraint $0 \leq I \leq Y/p$. Fortunately, the assumption $\xi'(H) < 0$ guarantees that H_t^* is unique. It can also be shown that $I^* < Y/p$ (see n. 12 below); however, the possibility that $I_t^* = 0$ cannot be ruled out.

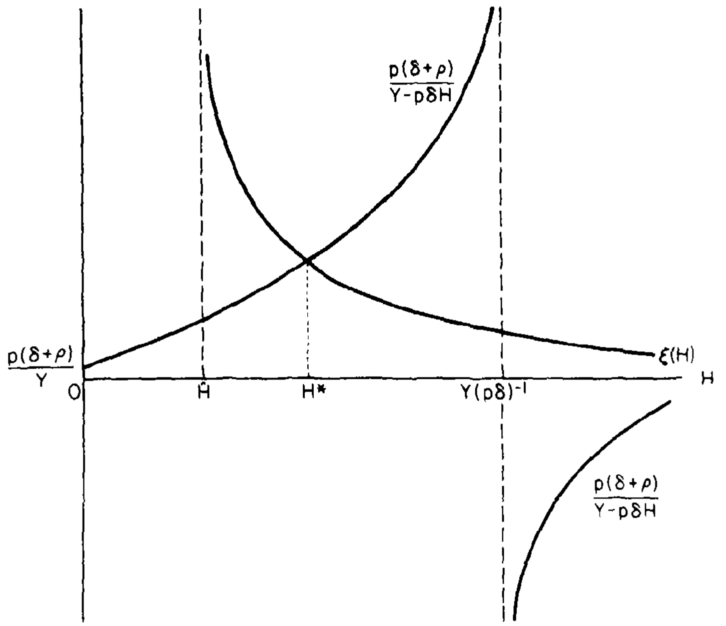


FIG. 2.—Determination of the optimal stock of health capital when death is endogenous

equation (24):[12]

$$\dot{H}^* = \frac{\delta p[H\xi(H) + 1]}{\xi'(H)(Y - p\delta H) - \xi(H)\delta p}. \tag{24}$$

As long as δ continues to increase over time, $p(\delta + p)(Y - p\delta H)^{-1}$ must eventually intersect $\xi(H)$ at $H = \hat{H}$, or, if $\xi(H)$ is asymptotic to $\hat{H}$, $\hat{H}$ must exceed $Y(\delta p)^{-1}$, implying that an optimal H^* does not exist. In either case, death occurs in finite time. It should also be noted that the proportion of time that the individual will spend ill increases as the individual ages, due to the fact that as the stock of health capital falls, the probability of being ill at any instant rises.

As indicated above, once the optimal stock of health capital is known, investment in health may be obtained by solving equation (3) for I^*,

$$I^* = \dot{H}^* + \delta H^*.^{13} \tag{25}$$

---

[12] An important result of figure 2 is that $I_t^* < Y/p$, which guarantees that the individual always has a positive level of consumption. As figure 2 implies, $\delta H_t^* < Y/p$; however, if $\dot{H}_t^* < 0$, $I_t^* < \delta H_t^*$, implying $I_t^* < Y/p$.

[13] Equation (25) holds, strictly speaking, only when $I > 0$. If $I = 0$, equations (23) and (24) no longer apply and the individual is said to be in a blocked interval. However, conditions may be placed on the path of δ through time to guarantee that the individual will never be in blocked interval, except at the time of death.

The behavior of I^* over the life cycle is then determined by differentiating (25) with respect to time,

$$I^* = \ddot{H}^* + \dot{\delta} H^* + \delta \dot{H}^*. \tag{26}$$

Unfortunately, previous assumptions no longer guarantee that I is unambiguously negative after some point in the life cycle, as was the case in IIB. The explanation for this is simple: Even though marginal benefits from increasing the stock of health capital, ψ, decline over time, replacement investment, δH_t, may increase as a result of $\dot{\delta} > 0$. Hence it is no longer necessarily the case that $I < 0$.

Before leaving figure 2 it is interesting to examine how the stock of health capital varies among individuals in different income classes and among individuals living in different environments. An immediate implication of the diagram is that among individuals with different incomes who are otherwise identical, wealthier individuals will maintain higher stocks of health capital than poorer persons. This implies, through the function $1 - F(H)$, that low-income individuals, ceteris paribus, will spend more time ill than their wealthier counterparts.

A less obvious implication of equation (23) is that persons who live in polluted environments will choose to maintain lower stocks of health capital and, for this reason, will be ill more often than individuals who live in healthier environments. One way to capture the effects of pollution is to say that health capital deteriorates faster, that is, that δ_t is higher, the more polluted the environment. This implies that pollution has a cumulative deleterious effect on health which, if not offset by investment in health, will cause individuals to be ill more often and die sooner in a high-pollution area. Since $\partial H_t^* / \partial \delta_t < 0$, the model shows that individuals will not fully offset the effects of pollution by investing in health, and hence that the optimal health stock will be lower the more polluted the environment.

Finally, the model suggests that individuals who are more likely to become ill will compensate for this fact by maintaining higher stocks of health capital. In the language of IIA, individuals who have a higher probability of being ill (for a given H) have distribution functions $F(\bar{H})$ which lie below those of individuals who are less likely to be ill. To see how this affects the optimal stock of health capital, let us consider the family of density functions

$$f(\bar{H}) = \begin{cases} \alpha e^{-\alpha \bar{H}}, & \bar{H} \geq 0 \\ 0, & \bar{H} < 0 \end{cases}, \qquad \alpha \geq 1.\text{[14]} \tag{27}$$

This family satisfies the condition $\xi'(H) < 0$, and has the intuitively appealing property that small values of $\bar{H}$ are more likely to occur than large values of $\hat{H}$. The corresponding distribution function is given by

---

[14] Without loss of generality, H is assumed to be measured as a deviation from $\bar{H}$.

$$F(H) = 1 - e^{-\alpha H}. \tag{28}$$

Since

$$\frac{\partial F(H)}{\partial \alpha} = He^{-\alpha H} \tag{29}$$

is positive for $H > 0$, an individual for whom α is small has a lower probability of being well for any H than a person for whom α is large.

The optimal health stock depends, however, on $\xi(H)$, the conditional probability density of illness. For the distribution function (28),

$$\xi(H) = \frac{\alpha e^{-\alpha H}}{1 - e^{-\alpha H}} \tag{30}$$

and

$$\frac{\partial \xi(H)}{\partial \alpha} = \frac{(1 - \alpha H)(1 - e^{-\alpha H})e^{-\alpha H} - \alpha He^{-\alpha H}}{(1 - e^{-\alpha H})^2}, \tag{31}$$

which is negative for $H \geq \alpha^{-1}$. Thus in terms of figure 2, $\xi(H)$ will be higher (except for small values of $H - \hat{H}$) the lower is α. This implies, in particular, that individuals who are more likely to be ill will maintain higher stocks of health capital.

III. Occupational Choice as Investment in Health

After analyzing what is conventionally meant by investment in health, let us now examine a more subtle way in which an individual affects his stock of health capital - namely, through choice of occupation. In this section the problem facing the individual is one of trading higher wages for greater exposure to a pollutant which has deleterious effects on the worker's health and increases the probability that he will die at an early age. This type of trade-off is especially important in industries such as coal mining and asbestos manufacturing; indeed, a major justification for the present model is the insights which it yields into the decision to work in such occupations.

A. The Model

In order to focus on the problem of occupational choice, we shall assume at each t that the individual has $\bar{L}$ of labor which he divides between a "healthy" occupation (farming) and an "unhealthy" occupation (coal mining), which pays a higher wage but exposes him to a potentially lethal pollutant. If L_t denotes the number of hours worked in the unhealthy occupation, then the individual's income at t may be written

$$Y_t = \bar{L} + wL_t, \tag{32}$$

where w represents the wage differential between the two occupations and the wage in the healthy occupation is taken without loss of generality to be one. As the individual works in the unhealthy occupation he is assumed to absorb a pollutant, such as coal dust, at a rate of b units per hour. The rate of change in the pollution stock is thus given by

$$\dot{P} = bL - \alpha P, \tag{33}$$

where α is the rate at which the stock decays.

Income and the pollution stock affect the individual's utility in the following way. If the individual is alive at time t, he receives utility from income according to the strictly concave and increasing function $U(Y)$, where $U(\cdot)$ has the properties listed in (4). If the individual is dead, income is zero and no utility is received. Although by assumption the pollutant has no immediate effect on utility, a sufficiently high concentration of the pollutant will cause death.[15] Formally, death occurs if $P \geq \bar{P}$, where $\bar{P}$ is the lethal dose of the pollutant. We shall assume that there exists a true $\bar{P}$ which does not change throughout the individual's lifetime but which is unknown to the individual. Subjective beliefs about $\bar{P}$ at time 0 are reflected in the probability density function $g(\bar{P})$, which is distributed over the interval (P_0, ∞) with $g(\bar{P}) > 0$.[16] As time passes, the knowledge that the individual has not yet died provides information which may be used to update the distribution $g(\bar{P})$. It is shown in Appendix B that if the worker revises his beliefs about $\bar{P}$ at some time τ, $\tau > 0$, the path of employment based on these updated beliefs will be consistent with the path of employment chosen at 0.

The above assumptions imply that the probability that the individual dies at t is $G(P_t)$, where $G(\cdot)$ is the cumulative distribution corresponding to $g(\bar{P})$. However, for the purpose of computing expected utility, it is not the probability that the individual dies at t but the probability that he is alive at t which is of interest. The probability that the individual is alive at t is computed in Appendix A by noting that the event "the individual is alive at t" is equivalent to the event "the individual does not die prior to t." As is shown below, the probability that the individual does not die prior to t is the probability that $\bar{P}$ exceeds the *largest* value of P in the interval $[0, t]$. Hence

$$P(\text{alive at } t) = \int_{\max\limits_{0 \leq \tau \leq t} P_\tau}^{\infty} g(\bar{P})\, d\bar{P} \equiv \Lambda\left(\max_{0 \leq \tau \leq t} P_\tau\right). \tag{34}$$

By controlling the size of his pollution stock the individual thus affects the probability that he will live or die and in turn determines his expected utility,

---

[15] Many industrial pollutants do, of course, have immediate side effects, which suggests that utility should depend on P as well as on Y. Unfortunately, the case where $U = U(Y, P)$ is intractable, for reasons that will become clear below.

[16] Since it is assumed that $\bar{P}$ remains constant throughout the individual's lifetime, the knowledge that the individual is alive at $t = 0$ implies that $P_0 < \bar{P}$.

$$EU_t = U(Y_t)\Lambda\left(\max_{0 \leq \tau \leq t} P_\tau\right). \tag{35}$$

The interesting question is how the relationship between the pollution stock and expected utility affects the individual's choice of occupation.

Formally, the problem of occupational choice consists of selecting a path L_t to maximize discounted expected utility

$$\int_0^T e^{-\rho t} U(Y)\Lambda\left(\max_{0 \leq \tau \leq t} P_\tau\right) dt \tag{36}$$

subject to (33) and to the inequality constraint $L \geq L \geq 0$. As equation (36) implies, the individual is assumed to die at T from "natural causes" if he does not die sooner as a result of his job.

The analysis of equation (36) is complicated by the fact that instantaneous expected utility depends not only on the current value of the pollution stock but on all previous values of P as well. However, the objective function may be simplified by showing that the optimal pollution stock never decreases, and hence that

$$\Lambda\left(\max_{0 \leq \tau \leq t} P_\tau\right) = \Lambda(P_t).$$

To see this, note that in equation (36) P affects utility only by altering the probability that the worker is alive. Since the probability that the worker is alive at t is the probability that $\bar{P}$ exceeds the largest value of P in the interval $[0, t]$, the individual can never increase his utility at t, or at any time following t, by setting $\dot{P}_t < 0$. Since this argument holds for any t, the optimal pollution stock can never decrease and

$$\Lambda\left(\max_{0 \leq \tau \leq t} P_\tau\right) = \Lambda(P_t). \tag{37}$$

With this simplification, equation (36) may be written

$$\int_0^T e^{-\rho t} \Lambda(P) U(L + wL)\, dt, \tag{38}$$

and the problem of occupational choice is seen to be analogous to the problem of investment in health studied in Section IIB.

B. Life-Cycle Behavior of Employment and Pollution

Note that in simplifying (36) an important result has been obtained, namely, that P_t never decreases. While intuition might suggest that an individual working in a risky occupation would build up his pollution stock when he was young, allow it to decrease later on in life, and possibly return to the risky occupation in old age, the model indicates that as long as the individual suffers no immediate discomfort from the pollutant, nothing is gained from allowing P to decrease since decreasing P cannot increase the probability that the individual will live. This suggests that individuals will take "recuperative breaks" from an unhealthy occupation only if the pollutant decreases their instantaneous utility. The fact that the job increases the probability of death is not, by itself, sufficient reason for leaving the occupation.

A related question of interest is how L_t, the path of employment in the risky occupation, varies over the life cycle. Two cases may be distinguished. If initially $P_0 = 0$ and $L_0 = 0$, then it may be shown that the individual will never work in the risky occupation.[17] On the other hand, if the wage is high enough that $L_0 > 0$, equation (33) and the condition $\dot{P} \geq 0$ imply that the individual will always work some amount of time in the risky occupation. The path of employment in the latter case may be derived from the results of Section IIB by noting that a high stock of pollution is equivalent to a low health stock, while an increase in investment in health ($\dot{I} > 0$) corresponds to a decrease in the number of hours worked in the risky occupation ($\dot{L} < 0$).[18] Since $\dot{P} \geq 0$ ($\dot{H} \leq 0$), only paths analogous to (2) and (4) in figure 1 are possible.[19] Thus employment in the risky occupation will begin at $L_0 > 0$, decrease as the individual grows older, and then increase as retirement (time T) approaches. This implies that an individual who chooses to work in an unhealthy occupation will expose himself to large amounts of pollution early in life, decreasing his exposure in middle age to prevent too rapid an accumulation of the pollutant. As the individual grows older, the negative effects of pollution last for a shorter length of time and employment in the risky occupation increases. Thus, even though the optimal pollution stock never decreases, the amount of time worked in the risky occupation does decrease during the life cycle to prevent P from increasing too rapidly.

---

[17] The condition which determines L as a function of λ and P is $wU'(Y)\Lambda(P) + b\lambda \leq 0$. Hence if $|wU'(L)/b| < -\lambda_0$, $L_0 = 0$. As long as $L_t = 0$, however, it can be shown that λ decreases over time and hence that $-\lambda_t$ increases in absolute value. Hence if $L_0 = 0$ and $P_0 = 0$, $L_t = 0$ thereafter.

[18] For a complete analysis of the solution to (38) the reader is referred to Cropper (1976).

[19] The fact that paths corresponding to (1) and (3) are no longer possible implies that there is an important difference between the models of Sections II and III. The difference is in the nature of the randomness in each model. In Section II the probability that the individual is well at t is, by assumption, independent of the probability that he is well at t'. Hence by setting $\dot{H} < 0$, the individual can always increase the probability of being well. In Section III, however, the probability of being alive at t is not independent of the probability of being alive at t'; in fact, $P(\text{alive at } t) \leq P(\text{alive at } t')$, where $t' < t$. Thus the individual can never increase the probability that he is alive at t by setting $\dot{P}_t < 0$. For this reason $\dot{P} \geq 0$, and one of the two strategies which is possible in Section II is ruled out in Section III.

In addition to the above conclusions, the present model has interesting implications for an important real-world problem. Frequently, it is learned that a job previously thought to be harmless actually exposes workers to a dangerous pollutant. An important question is how workers should optimally react to this information, assuming they have already ingested significant amounts of the pollutant but also assuming that they suffer no noticeable side effects from it. A commonly held view is that workers should leave the unhealthy occupation in an attempt to neutralize the effects of the pollutant. In many cases, however, workers are observed to remain in risky occupations after being warned of the dangers involved. The model suggests that this behavior is indeed optimal. By analogy with figure 1, when T is "small" the worker will at first reduce the number of hours worked in the unhealthy occupation; however, L will eventually increase, with the individual working full time in the risky occupation as retirement approaches. Thus the individual's failure to decrease his exposure to risk need not be the result of myopia or lack of employment opportunities but may reflect a rational decision.

IV. Conclusion

This paper has presented two models of investment in health which explicitly recognize the random nature of illness and death In the first model the demand for preventative medical care (checkups, dietary supplements, etc.) is derived from the more basic demand for health capital. Individuals desire to increase their stock of health capital in order to decrease the probability of illness. This is because an individual, when ill, receives no utility from consumption; hence investment in health increases expected utility of consumption. Against this increase in expected utility the individual must balance the cost of investing in health, which depends on the price of medical services, the rate of depreciation of health capital, and the individual's subjective discount rate.

The important question, of course, is how investment in health and the optimal health stock vary throughout the life cycle. In human-capital theory, individuals usually invest most heavily in themselves when they are young, with the amount of time devoted to training declining steadily thereafter (see Blinder and Weiss 1976). If the time of death were known with certainty, one would expect that investment in health would also be concentrated at the beginning of one's life and decrease as the time of death approached. If the conditional probability of illness is a decreasing function of the health stock, this is in fact the case. However, a major difference between investment in human capital and investment in health is that in a model of health capital the time of death cannot be treated as exogenous. This problem is remedied in Section II by assuming that death occurs if the individual is ill 100 percent of the time. If one also assumes that the rate of decay of health capital increases with age, then it can be shown that the individual will live a finite life and that his stock of health capital will decrease as the time of death approaches. This, in turn, implies that the fraction of time spent ill increases as the individual ages and that recuperative medical expenditures should, therefore, increase. Preventative medical expenditures, on the other hand, may decrease as the individual ages.

In the second model, investment in health is related to occupational choice. Casual observation suggests that for individuals with a fixed stock of human capital there is often a trade-off between earning higher wages and working in an unhealthy environment. This trade-off is especially important when the higher-paying job exposes the worker to an industrial pollutant such as asbestos, which in sufficiently high concentrations can cause death. In the model of Section III the individual must choose between a "safe" occupation which has no effects on health and a higher-paying occupation which exposes the worker to a potentially lethal pollutant. Increases in the stock of the pollutant increase the probability of death; hence the worker must trade the opportunity to earn a higher income against an increased probability of dying.

If it is assumed that the pollutant does not yield immediate disutility to the worker, it can be shown that the optimal pollution stock never decreases and that the worker if he ever works in the risky occupation will always expose himself to some amount of pollution. This implies, in particular, that an individual who has been working in an occupation which is discovered to be hazardous to the health need not quit the occupation once he is informed of the dangers to his health. Thus what upon casual observation may appear to be irrational behavior is shown by our model to be consistent with rational decision making.

Appendix A

To compute the probability that the individual is alive at time t, let us define the event, "the individual does not die at τ" and denote it $\bar{D}_\tau$:

$$\bar{D}_\tau = \{\bar{P} \mid \bar{P} \geq P_\tau\}. \tag{A1}$$

Since the probability that the individual is alive at t is the probability that he has not died at any time up to and including t,

$$P(\text{alive at } t) = P\left\{\bigcap_{0 \leq \tau \leq t} \bar{D}_\tau\right\}. \tag{A2}$$

Note, however, that because the lethal dose of the pollutant does not change throughout the individual's lifetime,

$$\bigcap_{0 \leq \tau \leq t} \bar{D}_\tau = \left\{\bar{P} \mid \bar{P} \geq \max_{0 \leq \tau \leq t} P_\tau\right\}. \tag{A3}$$

Thus

$$P\left(\bigcap_{0\le\tau\le t}\bar{D}_\tau\right) = \int_{\max\limits_{0\le\tau\le t} P_\tau}^{\infty} g(\bar{P})\,d\bar{P}. \tag{A4}$$

Hence if

$$\Lambda(P) \equiv \int_{P}^{\infty} g(\bar{P})\,d\bar{P},$$

$$P\left(\bigcap_{0\le\tau\le t} D_\tau\right) = \Lambda\left(\max_{0\le\tau\le t} P_\tau\right). \tag{A5}$$

Appendix B

The purpose of this Appendix is to show that the optimal path of employment in Section III allows for revision of expectations about $\bar{P}$, the lethal dose of the industrial pollutant. The method of proof follows Long (1975).

Let $g(\bar{P} \mid P_\tau)$ be the density function over $\bar{P}$, given that P_τ of the pollutant has already accumulated and that death has not yet occurred.

One may likewise define

$$\Lambda(P_t \mid P_\tau) \equiv \int_{P_t}^{\infty} g(\bar{P} \mid P_\tau)\,d\bar{P} \tag{B1}$$

to be the probability that the lethal dose is at least of size P_t, given that P_τ of the pollutant has accumulated to date.

The question is whether the path L_t, which maximizes

$$\int_0^T e^{-\rho t}\Lambda(P_t \mid 0)U(\bar{L} + wL)\,dt \tag{B2}$$

subject to

$$\begin{aligned} \dot{P} &= bL - \alpha P, \\ \bar{L} &\ge L \ge 0, \\ P_0 &\text{ given,} \end{aligned} \tag{B3}$$

is consistent with the path chosen at time t', provided death has not yet occurred, to maximize

$$\int_{t'}^{T} e^{-\rho(t'-t)}\Lambda(P_t \mid P_{t'})U(\bar{L} + wL)\,dt \tag{B4}$$

subject to (B3).

Since $\Lambda(P_t, P_{t'} \mid 0)$, the joint probability that $\bar{P} > P_t$ and $\bar{P} > P_{t'}$, equals $\Lambda(P_t \mid 0)$ for $P_t > P_{t'}$, expectations will be revised according to

$$\Lambda(P_t \mid P_{t'}) = \frac{\Lambda(P_t \mid 0)}{\Lambda(P_{t'} \mid 0)}. \tag{B5}$$

It thus follows that, given $P_{t'}$, the revised probability of being alive differs from the initial probability only by a multiplicative constant, and (B2) and (B4) will, therefore, yield identical solutions.

References

Blinder, Alan S., and Weiss, Yoram. "Human Capital and Labor Supply: A Synthesis." *J.P.E.* 84, no. 3 (June 1976): 449–72.

Cropper, M. L. "Two Models of Investment in Health." Working Paper no. 14, Univ. California, Riverside, 1976.

Grossman, Michael. "On the Concept of Health Capital and the Demand for Health." *J.P.E.* 80, no. 2 (March/April 1972): 223–55.

Long, Ngo Van. "Resource Extraction under the Uncertainty about Possible Nationalization." *J. Econ. Theory* 10 (February 1975): 42–53.

Thaler, Richard, and Rosen, Sherwin. "The Value of Saving a Life: Evidence from the Labor Market." In *Household Production and Consumption*, edited by Nestor E. Terleckyj. New York: Columbia Univ. Press (for Nat. Bur. Econ. Res.), 1975.

Inflationary Finance and Growth

Bijan B. Aghevli

International Monetary Fund

The optimal rate of monetary expansion is derived for the case when the government resorts to deficit financing to finance its development expenditure. It is argued that, while higher rates of monetary expansion increase investment and contribute to future consumption, the consequent inflation imposes welfare costs by reducing the level of real balances held by the public. A theoretical framework is developed and its empirical implications considered by deriving the optimal rate of monetary expansion which maximizes the discounted flow of total consumption—material consumption minus the disutility of holding suboptimal levels of real balances—over time.

I. Introduction

The policy of deficit financing has been repeatedly criticized by economists, most notably Martin Bailey (1956), Robert Mundell (1965), Alvin Marty (1967, 1973), and Milton Friedman (1971). The basic argument against deficit financing has been that the consequent inflation imposes a substantial welfare cost on the holders of real balances which can be measured by the area under the demand curve for the real balances. More recently, Charles Cathcart (1974) and Jacob Frenkel (1976) have looked into the dynamic aspects of the welfare cost of inflation when the inflation expectations adjust with a lag to the actual price developments.

In most of the literature concerned with inflationary finance, however, there is no explicit specification of the impact of the government expendi-

The work on this paper was in part carried out while the author was a research fellow on the International Monetary Research Programme at the London School of Economics. Grateful acknowledgment is given to Harry Johnson, whose comments on an earlier draft resulted in substantial improvements of the paper. The author is also grateful to George Borts, Michael Mussa, and Alexandre Swoboda for valuable comments and suggestions. The views expressed and all the remaining errors are the sole responsibility of the author.

[*Journal of Political Economy*, 1977, vol. 85, no. 6]
© 1977 by The University of Chicago. All rights reserved.

ture on the real sector. If one is prepared to argue that in many of the developing countries, due to an inadequate tax system, the major portion of the developmental programs has to be financed by an inflation tax, then the welfare cost of inflation should be directly compared to the benefits associated with these programs. Mundell (1965) has developed a model in which government receipts are used for production of capital goods, thereby raising the real rate of growth. Assuming a constant capital-output ratio, Mundell shows that maximum growth is achieved when the rate of monetary expansion is set to maximize the real government deficit. Mundell then argues that for plausible coefficients the maximum rate of growth achieved by deficit financing is rather small, while the welfare cost of inflation is quite large. This argument, however, overlooks the fact that even small increases in the rate of growth of output could be quite significant when the corresponding increase in the level of discounted consumption over time is considered.[1]

The purpose of this paper is to generalize the Mundell analysis in order to derive the optimal rate of monetary expansion for the case when the government resorts to deficit financing to generate additional investment. The analysis differs from the Mundell framework in some of its key elements. Following Johnson (1967), the utility of cash balances is viewed as a substitute for material consumption and is thus included in the definition of total consumption. The optimal rate of monetary expansion is then derived in order to maximize the discounted flow of total consumption - material consumption minus the disutility of holding suboptimal levels of real balances. The formal model is developed in Section II. The empirical implications of the model are considered in Section III, by substituting plausible values of various parameters of the model. It is shown that, under certain assumptions, a case can be made for a certain amount of deficit financing leading to moderate rates of inflation. Some dynamic aspects of the analysis are briefly considered in the concluding section.

II. A Model of Inflationary Finance and Growth

In order to incorporate explicitly the role of government in the production process, a simple neoclassical growth model is developed in which output per capita, y, is assumed to be an increasing function of the capital-labor ratio, k, according to the following Cobb-Douglas production function:

---

[1] Assuming a value of 3 for the inflation coefficient, the corresponding rate of monetary expansion derived by Mundell is 33.3 percent. For plausible values for velocity and capital-output ratio, Mundell shows that the corresponding rate of growth is 0.8 percent per annum. For a country with a 1 percent real growth rate (not an unlikely figure, if we consider some of the less fortunate developing countries), however, this implies an 80 percent rise in the growth rate itself, which is by no means trivial. This was first pointed out by Johnson (1967, p. 289).

$$y = f(k) = \beta_0 k^{\beta_1}; \qquad 1 > \beta_1 > 0. \qquad (1)$$

It is further assumed that private savings which are channeled into private investment are a constant fraction, s, of total output. While this relation might be regarded as too simplistic, it will serve as a good approximation for our purposes.[2] If the marginal propensity to save is very low, the steady-state level of consumption will fall considerably short of the Golden Rule level. The authorities might then wish to augment private capital formations if they believe that (1) the private rate of discount is greater than the social rate of discount[3] and (2), if due to imperfections in the capital market, the private rate of return on investment falls short of the social rate. The latter issue is probably of more importance, since the shortage of private savings and investment in the developing countries is for the most part due to imperfections in the capital markets. The planning authorities might find it more efficient to undertake the necessary capital expenditure than to engage in the long and slow process of removing bottlenecks in the capital market. Moreover, the governments might argue that they are in a better position to decide on the relative merits of different investment projects than is the private sector. Clearly, many of these issues are debatable and are the subject of intensive research in the literature.[4] My purpose here is to focus on the case in which the government is forced to resort to deficit financing as the main source of its capital expenditure. Some of the implications of these factors will be considered later in the paper.

Assuming that an inflation tax is used as the only source of government capital expenditure, the level of capital formation per head provided by the authorities is given by pm, where p denotes the rate of monetary expansion and m denotes the level of real balances per capita. Total capital formation per capita (private as well as government) is given by expression (2):

---

[2] For a critique of the simple Keynesian savings function, see Johnson (1967). In an alternative formulation proposed by Tobin (1965) and Johnson, the savings function is specified as a function of "disposable income" which includes the additions to the public's holdings of real balances as well as the value of output. In this formulation, a rise in the rate of monetary expansion would reduce total savings even though the level of private capital formation increased. That is, the increase in the rate of inflation would increase the public's desired ratio of monetary wealth to capital. As the private sector attempts to restore its desired portfolio of monetary to nonmonetary wealth, savings in the form of capital formation will increase. For a survey of these issues, see Dornbusch and Frenkel (1973).

[3] It could be argued that governments, compared with individuals, should give more weight to future generations' welfare, leading to a lower rate of social discount than the private rate. The issue of the appropriate rate of discount is subject to much debate in the growth literature (see Baumol 1968), and it will not be developed further here.

[4] For a discussion of some of these issues in a more practical context, see Little and Mirrles (1974).

$$sf(k) + \rho m. \tag{2}$$

Further, imperfections in the capital market are likely to add to the importance of money as a store of value. That is, given the limited availability of financial assets, the public holds most of its wealth in the form of money or nonproductive physical goods (i.e., land, jewelry, etc.). The opportunity cost of holding money is then captured mostly by the expected rate of inflation, π, as the relevant substitution is between money and real goods (as opposed to capital).[5] The demand for real balances per capita, m, can be written according to the following Cagan-type function:

$$m = f(k)\alpha_0 e^{-\alpha_1 \pi}, \qquad \alpha_0, \alpha_1 > 0. \tag{3}$$

The public in general will not fully anticipate the inflation rate in the short run as expectations of inflation adjust to the actual inflation with a lag. The divergence between the actual and anticipated inflation leads to various costs associated with the distribution of income and resources. Some of the implications of these costs will be discussed later, but for now it is assumed that inflation is fully anticipated, in order to avoid the complications associated with the adjustment of price expectations. In the steady state, however, the anticipated (as well as the actual) inflation is given by the excess of the rate of monetary expansion, ρ, over the exogenously given rate of population growth, λ.[6]

$$\pi = \rho - \lambda. \tag{4}$$

Using expressions (2), (3), and (4), the following relationship can be written, where a dot denotes the time derivative:

$$\dot{k} = s_e f(k) - \lambda k \tag{5}$$

where

$$s_e = s + \rho g(\rho)$$

and

$$g(\rho) = \alpha_0 e^{-\alpha_1(\rho - \lambda)}.$$

The term $g(\rho)$ is the inverse of velocity, and the term s_e can be thought of as the effective savings ratio, which can be altered by changing the rate of monetary expansion. As the rate of monetary expansion increases, the effective savings ratio rises up to the point where government revenue from the inflation tax is maximized (i.e., $\rho = 1/\alpha_1$), and then it begins to decline toward the private marginal propensity to save, s, as the real government revenue from the inflation tax approaches zero.

---

[5] If one allows for substitution between money and capital, the real rate of interest should also be considered as an argument for the demand for real balances, since its omission would lead to suboptimal holdings of real balances at zero rate of inflation (see Frenkel 1976).

[6] In the absence of technological improvement, the steady-state rate of growth of output is given by the rate of population growth. It should be noted that technological growth of the Harrod neutral type could easily be incorporated in the model by redefining labor in terms of "efficiency units."

Consumption per capita can be written as a function of the rate of monetary expansion and the capital-labor ratio.

$$c = f(k)[1 - s - \rho g(\rho)] = f(k)(1 - s_e). \tag{6}$$

By setting (5) equal to zero and substituting into (6), the steady-state consumption can be derived as a function of the steady-state capital-labor ratio. Consequently, the steady-state level of consumption can be written as a function of the rate of monetary expansion.[7]

$$c(\rho) = \beta_0 \left[\frac{\beta_0 s_e(\rho)}{\lambda}\right]^{\beta_1/(1-\beta_1)} [1 - s_e(\rho)]. \tag{7}$$

The analysis above shows how the authorities can increase the level of consumption per head by creating forced savings through deficit financing and spending the proceeds on capital formation. But this process can induce substantial rates of inflation, increasing the opportunity cost of holding money relative to goods and leading to a reduction of real balances held by the public. Insofar as real balances provide the public with services, the fall in real balances implies a reduction in the utility yield of these balances which should be accounted for. That is, if money is viewed as a form of consumer capital good yielding a flow of utility, then a proper definition of consumption should include the services of money as well as services of goods.[8] On the other hand, money can be viewed as a producers' capital good and put directly into the production function.[9] As Johnson (1969) has shown, the two procedures lead to qualitatively similar results. That is, in both cases a fall in real balances will reduce immediate consumption. In the latter case, the reduction in consumption is due to the fall in output, while in the former case it is due to the fall in utility yield of money which should be included in the appropriate definition of consumption. Thus in both cases the fall in immediate consumption due to the fall in real balances can be compared with the change in future consumption associated with the change in the rate of capital formation.

---

[7] Substituting eq. (1) into (5) and setting $k = 0$ gives

$$k = \left[\frac{\beta_0 s_e}{\lambda}\right]^{1/(1-\beta_1)}.$$

The expression of k above can be substituted into eq. (6), which gives eq. (7).

[8] The concept of regarding the utility of cash balances as a substitute for material consumption in a growth model was first introduced by Johnson (1967). Johnson regards the utility yield of real balances as well as the flow of cash balances as part of the disposable income which augments consumption and savings.

[9] See, e.g., Levhari and Patinkin (1968) and Johnson (1969).

While real balances have been included as an argument in the production function in many of the recent theoretical papers in the monetary-growth literature, efforts to isolate this factor empirically have been at best preliminary.[10] Thus, for the purposes of this paper, money is viewed as a consumer capital good-yielding utility which can be derived from knowledge of the demand for real balances—an area which has been investigated empirically quite extensively. Following the literature on welfare cost of inflation, notably Bailey (1956) and Marty (1967, 1973),[11] the utility loss of a reduction in real balances is derived by integrating the demand for real balances under the assumption that the marginal utility of consumption is constant. This will also allow a comparison of the results of this study with those of previous authors. Let q denote the loss in utility associated with a reduction of real balances from $m(0)$, corresponding to zero rate of inflation, to $m(\pi)$, corresponding to π rate of inflation. Using equations (3) and (4), the following expression is derived for q:

$$q = \int_{m(\pi)}^{m(0)} \pi \, dm = f(k)\left[\frac{\alpha_0}{\alpha_1} - \left(\rho - \lambda + \frac{1}{\alpha_1}\right) g(\rho)\right]. \tag{8}$$

Total utility can then be defined as the utility of consumption of goods minus the utility loss associated with the reduction in real balances. Assuming that the marginal utility of consumption is constant, the utility loss associated with the reduction of real balances can be thought of as a loss of substitute consumption goods.[12] Total utility, u, can then be written as the total consumption of goods, c, minus the forgone services of real balances, q. From equations (6) and (8), it follows that

$$u(k, \rho) = c - q = f(k)\psi(\rho), \tag{9}$$

where

$$\psi(\rho) = \left(1 - s - \frac{\alpha_0}{\alpha_1}\right) + \left(\frac{1}{\alpha_1} - \lambda\right) g(\rho).$$

---

[10] See Sinai and Stockes (1972) and the subsequent comments by Khan and Kouri (1975).

[11] For a detailed analysis of the welfare cost of inflation leading to this formulation, see Bailey (1956) and Marty (1967, 1973). It should be noted that some subtler issues pertaining to the adjustment behavior of expectations and real balances discussed in Auernheimer (1974), Cathcart (1974), and Frenkel (1975, 1976) are ignored in order to avoid some mathematical complexities. In principle they could be incorporated into the analysis.

[12] In a more general formulation, the marginal utility of consumption could depend on the welfare cost of inflation associated with the distortions in the money market. That is, the social cost of providing the consumption good need not be zero, which implies that the gain in utility due to a change in consumption would be given by the integral of the product of the change in consumption times the associated distortion. For the purposes of this model, these difficulties are disregarded by assuming that the marginal utility of consumption is constant. This will not be a bad approximation; see, e.g., Johnson (1967) and Hause (1975).

Here $\psi(\rho)$ can be thought of as the effective consumption ratio, where effective consumption is appropriately defined to include the consumption of goods as well as the utility loss of services of forgone real balances. It should be noted that the sum of effective savings and consumption ratios will fall short of unity for positive rates of inflation, since the effective consumption will be less than the consumption of real goods.

The utility function above should be maximized over time subject to the dynamic adjustment of capital-labor ratio given by equation (5). In order to avoid technical details, the analysis is confined to an examination of steady-state growth paths.

Assume that the authorities discount the future stream of effective consumption, u, at the social rate of discount, δ. Leaving aside the difficulties associated with the determination of the social rate of discount for now, this implies the following maximization process:

Maximize

$$\int_0^\infty u(k, \rho)e^{-\delta t}\, dt$$

subject to

$$k = s_e(\rho)f(k) - \lambda k.$$

The modified Golden Rule solution of the maximization process above can be shown to be determined by the solution of equation (10):[13]

$$\left(-\frac{\delta}{\lambda} + \beta_1 - 1\right)\frac{\psi'(\rho)}{\psi(\rho)} = \beta_1 \frac{s_e'(\rho)}{s_e(\rho)}. \tag{10}$$

It should be noted that the smaller the rate of discount, δ, the larger will be the optimal rate of monetary expansion. This result can be seen intuitively, since higher rates of monetary expansion diminish present effective consumption—higher government deficits reduce present consumption of goods in addition to imposing a higher welfare cost of inflation. On the other hand, higher rates of monetary expansion increase present investment and, up to a point, contribute to the future stream of

---

[13] The Pontryagin maximum principle can be applied by constructing the current-value Hamiltonian H when p denotes the dual-state variable to k.

$$H = u(k, \rho) + p\dot{k} = f(k)\psi(\rho) + \rho[s_e(\rho)f(k) - \lambda k].$$

The maximization conditions can be written as follows:

$$\frac{\partial H}{\partial \rho} = f(k)[\psi'(\rho) + ps_e'(\rho)] = 0,$$

$$\dot{p} = \delta p - \frac{\partial H}{\partial k} = (\delta + \lambda)p - f'(k)[\psi(\rho) + ps_e(\rho)],$$

$$k = \frac{\partial H}{\partial p} = s_e(\rho)f(k) - \lambda k.$$

Setting $\dot{k} = \dot{p} = 0$ and solving the dynamic system above gives eq. (10).

consumption. Higher rates of social discount, however, put relatively less weight on future benefits and more weight on the present costs of higher government deficits, leading to lower optimal rates of monetary expansion and inflation.

III. Some Empirical Implications of the Model

In order to draw some empirical implications from the analysis above, some numerical estimates are provided in this section. The following plausible values for various parameters of the model are chosen as rough orders of magnitude to approximate the sorts of values one would expect to find in those developing countries which resort to substantial deficit financing:

$$\alpha_0 = 0.25 \qquad \alpha_1 = 0.5;\ 1.0;\ 3.0 \qquad \lambda = 0.03$$

$$\beta_0 = 25 \qquad \beta_1 = 0.25 \qquad s = 0.05$$

$$\delta = 0;\ 0.05.$$

The parameter α_0 is the inverse of the velocity of circulation of money corresponding to zero inflation. Parameter β_0 is an arbitrarily chosen scale factor which measures units of output. Parameter α_1 measures the responsiveness of the demand for real balances to the anticipated rate of inflation. Cagan's (1956) estimate of α_1 for a set of hyperinflation countries is about 0.5. As Bailey (1956) points out, however, Cagan's estimate might be biased downward, since his sample consisted only of hyperinflation countries. Other estimates provided in the literature[14] range from zero to 3 (most estimates are less than unity), 3 being the value used by Mundell (1965) and consequently by Marty (1967). As will be shown later, the analysis is crucially dependent on the value of α_1. Thus, following the available empirical work, three alternative values of 0.5, 1, and 3 are considered for α_1. The β_1 is the coefficient of capital in the Cobb-Douglas production function (i.e., capital's share of the output), and λ is the exogenously given rate of population growth.

The private savings rate, s, is assumed to be 5 percent as a typical value for the developing countries. Assuming no forced savings, this would imply a rather high rate of private discount (i.e., 12 percent).[15] If one believes that this is the appropriate rate to be chosen by the authorities for the social discount rate, then it follows that no additional capital formation by the government should take place. On the other hand, if one believes that the low savings rate in the developing countries is more a reflection of inadequate capital markets than of a high private discount rate, the authorities should use a lower social discount rate, which implies that they must augment private investment. While there is much debate in the literature on the appropriate rate of discount to be used by the planning authorities,[16] there are no hard and fast rules to provide us with an estimate of this variable. A value of 5 percent was chosen as a likely value for the social discount rate, δ. This would imply that the mean period of the authorities' planning period is about 20 years. A zero rate of discount is also considered as an alternative to represent no discounting (i.e., the Golden Rule solution).

---

[14] See, e.g., a collection of studies in Meiselman (1970) for Chile (Deaver), Argentina (Diz), and Korea and Brazil (Campbell). See also a cross-section study for 47 countries by Vogel (1974), a study of Indonesian inflation by Aghevli (1977), as well as the study by Aghevli and Khan (1977). Based on these studies on the demand for money for developing countries, the estimated value of α_1 ranges between zero and 3 where most of the estimates are below unity.

The optimal rate of monetary expansion is derived by solving equation (10) numerically. Given the optimal rate of monetary expansion and thus the effective savings ratio (i.e., $s + \rho g$), the steady-state capital-labor ratio and the corresponding level of output are given by equations (5) and (1). The level of goods consumption and the welfare cost of induced inflation can then be calculated from equations (6) and (8), respectively. The results are summarized in table 1, where optimal rates of monetary expansion and the corresponding steady-state levels of output, goods consumption, and the welfare cost of inflation are given in columns 1, 2, 3, and 4, respectively, for the range of values of α_1 and δ. The last column represents the ratio of the welfare cost of inflation to the level of real government revenue from deficit financing. This ratio can be thought of as the ratio of "collection cost" of the inflation tax to the amount collected. The last row of table 1 corresponds to the case when the rate of monetary expansion is set equal to the rate of population growth in order to eliminate inflation altogether.

As the results in table 1 indicate, the optimal rate of monetary expansion is crucially dependent on the value of α_1. For a given rate of monetary

---

[15] The modified Golden Rule solution of the system when there is no forced savings is given by the following relationship:

$$f'(k) = \lambda + \delta.$$

Utilizing the steady-state relation given by eq. (5) and employing the properties of the Cobb-Douglas production function yields

$$f'(k) = \frac{\beta_1 \lambda}{s}.$$

Equating the relationships above and using the assumed values for λ and β_1 implies that a savings ratio of 5 percent is consistent with a 12 percent rate of private discount.

[16] See nn. 3 and 4 above.

TABLE 1

OPTIMAL RATES OF MONETARY EXPANSION

	Rate of Monetary Expansion ρ (1)	Output $f(k)$ (2)	Con-sumption c (3)	Welfare Cost q (4)	Ratio of Welfare Cost to Inflation Tax $\frac{q}{\rho g(f(k))}$ (5)
$\alpha_1 = 0.5$:					
$\delta = 0$	0.70	132	109	2.97	0.18
$\delta = 0.05$	0.18	106	96	0.13	0.03
$\alpha_1 = 1.0$:					
$\delta = 0$	0.55	120	104	2.89	0.29
$\delta = 0.05$	0.15	100	92	0.12	0.06
$\alpha_1 = 3.0$:					
$\delta = 0$	0.25	102	94	1.21	0.37
$\delta = 0.05$	0.11	98	91	0.14	0.09
$\pi = 0$.	0.03	91	86	0	0

NOTE.—$\alpha_0 = 0.25$; $\beta_1 = 0.25$; $s = 0.05$; $\beta_0 = 25$; and $\lambda = 0.03$.

expansion, lower values of α_1 will allow the government to extract more forced savings from the public, leading to higher levels of steady-state output. Friedman (1971) has in fact used this point to argue that the maximum resources made available through deficit financing are quite small; he uses the alternative values of 2, 10, and 20 for α_1. Moreover, stating his impression that a value of 2 for α_1 "is well below the value that actually prevails in most developing countries" (p. 853), Friedman argues that, for values of parameters which are not at all implausible, a zero rate of inflation or even a negative one may provide the government with maximum revenue from the inflation tax. As already pointed out, however, most of the available empirical estimates of α_1 for the developing countries fall below unity. Thus in general Friedman's conclusions, which are based on very high (i.e., 10 and 20) values of α_1, will not hold.[17]

The results are also quite sensitive to the value of the social discount rate used by the authorities to discount future streams of consumption. This is a familiar problem facing those involved in cost-benefit analysis of development projects. While our choice of the social discount rate has been rather arbitrary, any formal derivation of this rate is outside the scope of the present analysis.

So far, it has been assumed that there is no possibility of increasing other forms of taxation. However, this assumption is quite unrealistic, as most countries can increase taxes to some extent, in spite of the large "collection costs" involved. If one had a reasonable estimate of these

---

[17] It is interesting to note that a high value of the inflation coefficient, α_1, implies a high interest elasticity of the demand for money, which is inconsistent with Friedman's other assertions.

"collection costs," other forms of taxation could be formally introduced into the model. It would then follow that the government should resort to deficit financing, as long as the ratio of the welfare cost of inflation to the inflation tax (given in table 1, col. 5) is less than the ratio of the collection costs of other taxes to the amount collected. But in the absence of any detailed information on the tax structure of the developing countries, only rough comparisons can be made. Bailey (1956) suggests that the collection costs of other forms of taxation, including welfare costs, compliance costs, and direct administrative costs for countries with relatively poor tax administration and tax morality are around 7 percent of the amount collected. Assuming that the existing level of taxation, which is normally associated with general government consumption, is already close to the maximum obtainable level, any further attempt to increase taxes will raise the collection costs to a much higher figure. On the other hand, for moderate rates of inflation, the ratio of the welfare cost of inflation to the inflation tax is shown to be quite small, particularly for the lower values of the inflation coefficient, α_1 (i.e., zero to unity), which are most likely to prevail in the developing economies.

IV. Conclusion

In this paper a model of the optimal rate of monetary expansion (and thus inflation) was constructed for those countries which resort to deficit financing in order to supplement private capital formation. It was shown that, when costs of inflationary finance (induced welfare costs of inflation) as well as the corresponding benefits (higher levels of consumption associated with higher rates of capital formation) are introduced into a growth model explicitly, one can make a case for moderate rates of inflation. Moreover, when future consumption is discounted at 5 percent, the authorities can reach the modified Golden Rule state in which the welfare cost of induced inflation is likely to be less than the costs involved in collecting other taxes. For lower rates of social discount, however, the ratio of the welfare cost to the inflation tax rises substantially if the inflation tax is used as the only form of taxation to reach the modified Golden Rule state. In this case, it will not be realistic to exclude the possibility of other forms of taxation for raising the necessary revenue.

The analysis is subject to a few more important qualifications. The assumption of a constant private savings rate might be questionable (see Johnson 1967), as it leads to the conclusion that all seigniorage and inflation-tax revenue comes at the expense of private consumption without affecting private savings. In a more general specification, the effects of monetary expansion on private savings could also be incorporated along the lines suggested by Tobin (1965) and Johnson (1967). Moreover, insofar as the analysis concentrates on examining mainly the steady-state solution of the model, some important dynamic considerations of the adjustment path of prices are neglected. As has been shown by Frenkel (1976), since the price expectations adjust with a lag to the actual price movements, the welfare cost of inflation could be overstated if one assumes (as is done in this paper) that inflation expectations are formed instantaneously. Also, as long as the inflation is not fully anticipated, there will be additional costs associated with the distribution of income and resources which are not captured in the framework employed here. Furthermore, the assumption that government spending is solely for productive investment will not in general be satisfied, since part of the increase in government spending may take the form of nonproductive consumption.

In light of the foregoing qualifications, the implications of the model for the attractiveness of even moderate amounts of deficit financing should be considered cautiously. Nevertheless, the basic framework captures some of the important issues in this area, and it can be generalized to include additional real-world complications.

References

Aghevli, Bijan B. "Money, Prices and the Balance of Payments: Indonesia, 1968–73." *J. Development Studies* 13, no. 2 (January 1977): 37–57.

Aghevli, Bijan B., and Khan, M. S. "Inflation Finance and the Dynamics of Inflation." *A.E.R.* 67 (June 1977): 390–403.

Auernheimer, Leonardo. "The Honest Government's Guide to the Revenue from the Creation of Money." *J.P.E.* 82, no. 3 (May/June 1974): 598–606.

Bailey, Martin J. "The Welfare Cost of Inflationary Finance." *J.P.E.* 64, no. 2 (April 1956): 93–110.

Baumol, William J. "On the Social Rate of Discount." *A.E.R.* 58 (September 1968): 788–802.

Cagan, Philip. "The Monetary Dynamics of Hyperinflation." In *Studies in the Quantity Theory of Money*, edited by Milton Friedman. Chicago: Univ. Chicago Press, 1956.

Cathcart, Charles D. "Monetary Dynamics, Growth and the Efficiency of Inflationary Finance." *J. Money, Credit and Banking* 6 (May 1974): 169–90.

Dornbusch, Rudiger, and Frenkel, Jacob A. "Inflation and Growth." *J. Money, Credit and Banking* 1, no. 1 (February 1973): 846–56.

Frenkel, Jacob A. "Inflation and the Formation of Expectations." *J. Monetary Econ.* 1, no. 4 (October 1975): 403–21.

———. "Some Aspects of the Welfare Cost of Inflationary Finance." In *Money and Finance in Economic Growth and Development: Essays in Honor of E. S. Shaw*, edited by R. I. McKinnon. New York, Marcel Dekker, 1976.

Friedman, Milton. "Government Revenue from Inflation." *J.P.E.* 79, no. 4 (July/August 1971): 846–56.

Hause, John C. "The Theory of Welfare Cost Measurement." *J.P.E.* 83, no. 6 (December 1975): 1145–82.

Johnson, Harry G. *Essays in Monetary Economics*, chap. 4. Cambridge, Mass.: Harvard Univ. Press, 1967.

———. "Inside Money, Outside Money, Income, Wealth, and Welfare in

Monetary Theory." *J. Money, Credit and Banking* 1, no. 1 (February 1969): 30–45.

Khan, Mohsin S., and Kouri, Penti. "Real Money Balances as a Factor of Production: A Comment." *Rev. Econ. and Statis.* 57 (May 1975): 244–45.

Levhari, David, and Patinkin, Don. "The Role of Money in a Simple Growth Model." *A.E.R.* 58 (September 1968): 713–53.

Little, I. M. D., and Mirrles, J. A. *Project Appraisal and Planning for Developing Countries.* London: Heinemann, 1974.

Marty, Alvin L. "Growth and the Welfare Cost of Inflationary Finance." *J.P.E.* 75, no. 1 (February 1967): 71–76.

———. "Growth, Satiety, and the Tax Revenue from Money Creation." *J.P.E.* 81, no. 5 (September/October 1973): 1136–52.

Meiselman, David, ed. *Varieties of Monetary Experience.* Chicago: Univ. Chicago Press, 1970.

Mundell, Robert A. "Growth, Stability and Inflationary Finance." *J.P.E.* 73, no. 2 (April 1965): 97–109.

Sinai, Allen, and Stockes, Houston H. "Real Money Balances: An Omitted Variable from the Production Function?" *Rev. Econ. and Statis.* 54 (August 1972): 290–96.

Tobin, James. "Money and Economic Growth." *Econometrica* 33 (October 1965): 671–84.

Vogel, Robert C. "The Dynamics of Inflation in Latin America." *A.E.R.* 64 (March 1974): 102–14.

Erratum

The following note was omitted from figure 1, page 782, in the paper by Isaac Ehrlich entitled "Capital Punishment and Deterrence: Some Further Thoughts and Additional Evidence," which was published in August 1977, vol. 85, no. 4, pages 741–88: "Arrows indicate approximate 95 percent and 99 percent confidence intervals for λ. Panels A and B depict the log-likelihood functions associated with transformations of the linear models used to estimate equation 1 in table 3 and equation (set) 2 in table 4A, respectively."

Book Reviews

The Theory of Equilibrium Growth. By A. K. Dixit.
London: Oxford University Press, 1976. Pp. vii+204. $12.00 (cloth); $5.00 (paper).

Dixit has written a useful book on the theory of equilibrium growth as developed by economists working within the tradition of neoclassical economic analysis. The assumptions are carefully stated (although the lack of essential institutional detail in their analyses is glossed over), and the various models are developed in a clear, concise manner. Dixit has succeeded in his aim of writing a book which bridges "the wide gap that exists between elementary textbooks on growth theory, such as Solow's *Growth Theory: An Exposition*, and the advanced ones like Wan's *Economic Growth* and Burmeister and Dobell's *Mathematical Theories of Economic Growth*" (p. v). The book is laid out in a manner which makes self-study by the student possible. It is well written and brief notes at the end of chapters help to guide the student's further reading in the subject matter covered. Dixit's use of cost and revenue functions instead of production functions permits considerable simplification in the development of two-sector models, and this makes his examination of these models much easier to follow than in other books.

The book begins with an examination of the method of equilibrium growth theory and production functions which treat capital as a factor of production on an equal footing with labor. A series of models are then presented. The textbook "Harrod-Domar" model (which should be distinguished more clearly from Harrod's model) is examined first, followed by the Solow-Swan model. Commendable warnings are given to the reader not to confuse analyses which are based on comparisons of equilibrium positions with those which are concerned with transitions. Sensible warnings of this kind are to be found throughout the book. For example, in discussing optimum saving, he notes that "considerations of an overlapping sequence of generations put a great strain on the intertemporal equilibrium story. Futures markets are even harder to visualize when not all agents in the economy exist at any common date" (p. 114). As its title indicates, however, the book concentrates on equilibrium growth. Some of the difficulties faced by neoclassical growth theory in its treatment of capital, as highlighted by the recent Cambridge controversies in capital theory (Harcourt 1972), are dealt with only parenthetically. Dixit refers his readers to the examination of such matters by Bliss (1975).

Separate chapters are devoted to the analysis of technological progress, optimum saving, two-sector models, natural-resource constraints, and disequilibrium models. The last is the briefest and least satisfactory chapter. Dixit's presentation of the Kahn-Robinson discussion of equilibrium growth is faulty because he reverses the order of causation and he ignores their emphasis on the importance of expectations. The desired rate of accumulation is a function of the expected rate of profit, while the actual rate of profit is determined by the rate of accumulation. It is thus misleading to "make the warranted growth rate an increasing function of the profit rate" (p. 177); the relationship goes the other way. If the rate of accumulation is below the warranted rate, then the Kahn-Robinson warranted growth rate is stable, given the resulting profits, when the expected rate of return on investment is such that firms increase the rate of accumulation (and conversely, if the rate of accumulation is above the warranted rate). Stability in this model does not depend, as argued by Dixit, on the amount of "saving relative to the accumulation that producers are willing to undertake" (p. 178). The amount of saving is determined by the rate of accumulation.

The book's strength lies in its careful presentation of neoclassical equilibrium growth theory; it is thus not surprising to find that investment is determined by saving, institutional features of the economy being examined are not stated, social relations are nowhere mentioned, perfectly competitive markets are the rule, and full employment prevails. Entrepreneurs have no role or income in equilibrium; the ownership of machines provides returns automatically in precisely the same way as the holding of a bond (p. 38). Distribution can thus be explained by marginal productivity without any residual appearing to complicate matters. Dixit is clearly aware of the difficulties in trying to present equilibrium growth theory as a reasonable description of a dynamic economy. The main justification he gives for studying this theory is that it provides important insights into the efficient intertemporal allocation of resources. "The equilibrium prices are then to be interpreted as planners' prices, or as imputed or shadow prices" (p. 11). He does, however, also claim "that study of equilibrium growth has an important role to play as a first approximation, or as a first step in a chain of reasoning" even though he states that he does not have faith in the "strong version of the 'neoclassical synthesis' " (p. 10). This claim is, of course, also based on faith. Given the cavalier treatment of historical time, uncertainty, expectations and investment behavior, and market imperfections, faith is needed to see it even as a useful "first step."

A. Asimakopulos

McGill University

References

Bliss, C. J. *Capital Theory and the Distribution of Income.* Amsterdam: North-Holland, 1975.

Harcourt, G. C. *Some Cambridge Controversies in the Theory of Capital.* Cambridge: Cambridge Univ. Press, 1972.

The Just Economy. By J. E. Meade.
Albany: State University of New York Press, 1976. Pp. xi+247. $17.50.

In some respects this book, which is the fourth volume in his series, Principles of Political Economy, is what we might expect from Meade: the lucid prose, the succinct presentation, the copious use of examples, and, of course, the lack of references. In some respects, however, I found this book rather disappointing. In this volume, Meade essentially takes equity decisions as given, while the efficiency effects of their implementation are often postponed until the next volume. This does not leave too much in between, and consequently a good deal of the book consists of descriptions of the issues, a look at some of the difficulties that would be encountered in a complete analysis, and a frustrating "tune into the next volume" as a coda. Also, part of the fun in reading Meade is encountering a new twist or insight among the familiar, but freshly presented, established material. Perhaps the gems are too well hidden for me (I never was that good at Easter egg hunts) or perhaps it is the scope of this particular volume, but I found the book to be disappointing in this regard. Much of it is a rather cursory, and not very inspiring, description of the conventional aspects of introducing equity considerations into economic models. Lest the reader get the wrong impression, I hasten to add that Meade, even when disappointing, is good and that this book could be read profitably by just about anybody—even those confirmed logical positivists to whom expressions like "the theory of justice" are an outright annoyance.

In chapter 2, we cross the Rubicon of welfare economics—the making of interpersonal utility comparisons. The crossing is easy since we are faced with the existence of policymakers who are engaged in the process of making such comparisons, and Meade feels we might as well assume they act "as if" they are making them in terms of a well-defined social welfare function (SWF). In chapter 3, the shortcomings of the compensation principle are discussed (at least, up to what is known [elsewhere] as the "Scitovsky paradox"). A formal analysis of separating the efficiency and equity considerations of any policy involving differential changes is also presented.

In chapter 4, Meade tackles some fundamental problems. First, the form of the SWF where we are introduced to the simplified two-person lexical maximin form. Meade is at his best here where he illustrates the equity implications of this egalitarian function with a rather devastating example. Curiously, this is one of the few places in the book where Meade's inspirational debt to Rawls, proclaimed in the preface, is apparent. While the unweighted utilitarian SWF is dispatched punctiliously in the same chapter, the bulk of the book is distinctly modified utilitarian in outlook since much of the formal analysis involves differential changes and the (non)constancy of the finite distributive weights is essentially irrelevant.

In the same chapter, Meade deals with the difficult problem of measuring individual welfare. As in his earlier volumes he adopts, rather uncritically I believe, current consumption as the appropriate measure. To be sure, Meade discusses some of the finicky problems associated with this welfare measure, namely, those involving differences in tastes or current circumstances not included in measured consumption, but the overall suitability of current consumption is not really challenged. I believe this works to the detriment of his discussion of the intertemporal aspects of equity in chapter 6. Amid an otherwise enlightening chapter which ranges from Hotelling's rule for resource depletion (eq. [6.3] on p. 87) to differences in public and private attitudes to uncertainty, Meade presents a simple model of four overlapping generations in an economy with growing output per head. Because the elasticity of the marginal utility of consumption is small, the lifetime consumption stream of each generation rises sufficiently rapidly so that the current consumption of older generations exceeds that of younger at the times the generations overlap, despite the rise in output per head over time. The fact that Meade concludes from this example that intergenerational equity requires redistribution from the older (but, over their lifetimes, poorer) generations to the younger is a clear indication of his commitment to an unsuitable measure of individual welfare rather than peculiar equity judgments on his part.

Chapter 7 is a fine discussion of the problems involved with measuring the inequality of the income distribution. While the text concerns itself with conventional indexes such as the Gini ratio, an attached note deals with indexes derived from social welfare functions such as Atkinson's index. In chapter 8, Meade discusses the degree of equality in the income distribution arising from competitive markets. Not surprisingly, it depends on the equality of factor endowments, the determination of which is examined in chapter 9. This chapter consists of a rather tedious example designed to illustrate the forces which perpetuate and reinforce existing inequalities in factor endowments through the intergenerational transmission of human and nonhuman capital.

Chapter 13 is an interesting chapter which catalogs the battery of income redistributive tools available to the government. A brief description and statement of the efficiency pros and cons of each are given. Anyone hoping to get a few insights from Meade about the optimal income tax literature or other such matters will be disappointed, since a detailed examination of the efficiency implications of these policies is postponed until the next volume. While this chapter is among the more interesting in the book, I find it unsatisfactory overall. Some of Meade's contentions are a little facile. For example, he argues that nationalization of industry as a distributive policy can yield efficiency gains because the income from state-owned industries can be used to reduce distortionary taxation. A careful reading indicates the true source of the efficiency gain. The properties are nationalized by means of lump-sum taxes - thus, the gains actually result from the indirect replacement of distortionary taxes by nondistortionary taxes. The appendix to this chapter is classic Meade. The social welfare effect of a given policy change is found in an uncertain, dynamic economy and is expressed as a page-long algebraic equation which is subsequently dissected and each term explained intuitively.

NEIL BRUCE

Queen's University, Kingston

Eurocurrencies and the International Monetary System. Edited by CARL H. STEM, JOHN H. MAKIN, and DENNIS E. LOGUE.
Washington, D.C.: American Enterprise Institute for Public Policy Research, 1976. Pp. 413. $12.00 (cloth); $5.00 (paper).

The international monetary system has changed substantially in recent years as the Bretton Woods system has been replaced by generally floating exchange rates. Parallel to this development there has been rapid growth and expansion of Eurocurrency markets. Although this parallelism does not necessarily indicate a causal relationship, the possibility that it does is intriguing and important enough to warrant serious research. *Eurocurrencies and the International Monetary System* contains the results of such research. The book is the outcome of a conference sponsored jointly by the American Enterprise Institute and the U.S. Treasury and held in Washington in October 1974. The purpose of the conference was to "shed light on the role of the Eurocurrency system in affecting the workings of the international monetary system and thereby the problems faced by the national financial systems" (p. 1).

After an introductory essay by John Makin, which also serves as a summary of what follows, the book is divided into four main sections, each corresponding to a conference session. Each part is in turn composed of a major paper (two papers in part 2), a set of critical comments from four to five official discussants, and an edited version of the general discussion among the conference participants. While including the comments of the discussants in a conference volume is useful, since they often provide interesting perspectives, most of the general discussion is not of the sort that makes it worthwhile printing, even in an edited form. This is not because the unofficial comments always lack interest but rather that they tend to be very long relative to the importance of the content. To reduce the size of an already long book it would therefore have been desirable to give only the highlights of the discussions.

Part 1 of the book is introduced by a paper entitled "Eurocurrencies and the Evolution of the International Monetary System" prepared by John Makin. In this paper Professor Makin attempts first to provide an explanation of the emergence of the U.S. dollar as an international money based on a theory of the evolution of a dominant monetary asset in a national economy. Makin then goes on to argue that the consequent increase in international financial integration led to less and less scope for autonomous national monetary policy, the unacceptability of which resulted in the breakup of the Bretton Woods system and the advent of generally flexible exchange rates. The Eurocurrency markets are supposed to have played a major role in this scenario by bringing about a degree of capital mobility which rendered national monetary policy unable to affect domestic economic activity.

After reviewing arguments relating the demand for international reserves to the introduction of flexible exchange rates, Makin concludes that there is likely to have been some reduction in the demand for international reserves and suggests that this may have been a reason for part of the inflationary surge in 1974. As to the effect on real trade flows, Makin argues that the effect of more flexibility of exchange rates is not unambiguous and he presents empirical evidence showing no apparent relationship between these variables. A final section of the paper deals with the question of the optimal degree of exchange-rate flexibility.

The official discussants of Makin's paper, all worth reading, take the author to task on a number of issues such as (*a*) the dating of the major shock to the Bretton-Woods system at 1971 with the abandonment of a passive balance of payments policy of the United States, (*b*) the implicit assumption of low goods market integration throughout the 1960s in Makin's discussion of the "overvalued" dollar, (*c*) the assumed critical role of the Eurodollar market in the breakdown of fixed exchange rates, and (*d*) the neglect of industrial organization aspects as reasons for the development of Eurocurrency market.

The topic of Part 2 of the volume is the efficiency of the international financial system and the appropriate way to model the mechanisms linking national financial markets together. A paper by Dennis Logue, Michael Salant, and Richard Sweeney entitled "International Integration of Financial Markets: Survey, Synthesis, and Results" is directed at the first of these questions. It starts out with a useful survey of different concepts of international integration of financial markets and reviews previous efforts at measuring them empirically. A very interesting finding of all these studies is the tendency for integration to have become stronger throughout the 1950s and the first half of the 1960s. After about 1967 no further integration can be detected, however, a conclusion which does not seem consistent with Makin's emphasis on the Eurodollar market as a reason for increased capital mobility.

Logue, Salant, and Sweeney extend current concepts of integration by applying the efficient market model to the international capital market. This model leads them to suggest covariability of interest rates as a criterion for integration. They then investigate short-term interest rates in seven countries by means of principal component analysis and conclude that these countries' financial markets are highly integrated since a single factor can explain over 80 percent of the generalized variance of these seven interest rates. As the authors themselves recognize, their empirical findings and method is not a test of the efficient market model, but they argue that their results show strong financial integration between the countries studies.[1]

The question of how to model interest linkages between national money markets is dealt with in the paper by Richard Herring and Richard Marston called "The Forward Market and Interest Rates in the Eurocurrency and National Money Markets." The authors present a theoretical model designed to examine the interaction between Eurocurrency deposit rates, forward exchange rates, and national interest rates. Their treatment represents an improvement over traditional models of forward exchange markets which typically assume that interest rates are exogenously determined, but the applicability to current events is seriously hampered by being limited to a regime of fixed exchange rates.

"The Eurocurrency Market, Exchange-Rate Systems, and National Financial Policies" is the title of the paper by Thomas Willett which served as a background for Part 3 of the conference. The reference to the Eurocurrency market is, however, minimal, being limited to a brief analysis of the inflationary impact of Eurocurrency growth. Apart from that, "the development of the Eurocurrency market is treated as a general increase in interest-sensitive capital mobility" (p. 195) despite the evidence surveyed by Logue, Salant, and Sweeney indicating no such relationship.

The main part of Professor Willett's paper is concerned with the effectiveness of monetary and fiscal policy under fixed and flexible exchange rates. Special attention is given to the importance of the degree of capital mobility. While a survey of the main theorems on policy effectiveness in open economies may be useful as a background for further analysis and discussion, the results are too well known to constitute a major part of a book of this kind. Built as they are on the Keynesian open economy model, the results could usefully have been scrutinized as to their robustness in face of relaxing the assumptions of fixed prices and wages on the one hand and a perfectly elastic supply of output on the other. It is not clear that either of these assumptions comes close to describing the current state of affairs in any economy of interest. If not, and if the theorems Willett reviews are very sensitive to these assumptions, the relevance of his discussion for the conduct of policy may be nil. A related criticism of the model on which Willett relies concerns the time period of the analysis. In the long run, neither fiscal nor monetary policy is likely to have an appreciable effect in a small open economy.[2] Arguments about stabilization policies in these economies must therefore relate to the dynamics of adjustment, and as such they ought to be couched in a model which makes this clear.

Three annexes are attached to Willett's paper. Of them, the first is the most interesting as it contains empirical evidence showing virtually no relationship between exchange rates and interest rates during the period February 1973–May 1974. These results thus cast "doubt upon the assertion that interest-rate changes have been a dominant determinant of exchange-rate behavior under the current float" (p. 223) and as such they indicate the need for alternative models and explanations.

The only paper which is devoted substantially to the Eurocurrency market is "Some Eurocurrency Problems: Credit Expansion, the Regulatory Framework, Liquidity, and Petrodollars" by Carl Stem. Professor Stem begins by enumerating the various developments and circumstances that have led to calls for controls over the expansion of the Eurocurrency market. He then raises the question as to what type of model is appropriate as a description of the expansion of Eurodollar deposits. He criticizes various versions of the multiplier approach mainly because of the difficulty of finding an adequate definition of the base upon which the expansionary process is built. Stem concludes that the nonbank financial intermediary credit expansion model is appropriate for analysis of the Eurocurrency system.

---

[1] It is worth pointing out here, since it is relevant for Makin's paper and for Willett's (see below), that the present reviewer has found an equally high degree of integration among goods markets in 21 OECD countries (see Genberg 1977).

[2] On both real and nominal variables under fixed exchange rates. With a flexible exchange rate, monetary policy of course controls the price level in the long run.

Among the possible ways to control the expansion of the Eurodollar market Stem analyzes and rejects both a system of reserve requirements imposed on Eurobanks and open market operations with the U.S. Federal Reserve acting as the world central bank. He instead argues convincingly that deregulation of U.S. banks is likely to be the most effective in limiting the growth of the Eurocurrency market. This conclusion is based on an analysis of the effect on the competitiveness of U.S. banks of current regulations limiting the payment of interest on certain kinds of deposits. Stem concludes his interesting paper with a section on the question of a lender of last resort for Eurobanks and an analysis of the impact of petrodollars on the Eurocurrency system.

As should be clear from the above, the book under review does not, except for Stem's paper, deal to any great extent with the Eurocurrency market. Instead its main emphasis is on aspects of capital mobility and fixed versus flexible exchange rates. Within these topics its strength is not in providing new results and conclusions but rather in surveying a large number of issues. As such it tends to be lengthy at times although occasionally an interesting hypothesis and suggestion for further research is provided. If the editors had attempted to reduce the length of the book fairly substantially, they would not have risked a reader skimming over these bright spots.

The book does not contain an index but has a very useful 37-page bibliography of which 51 entries are annotated.

HANS GENBERG

Graduate Institute of International Studies, Geneva

Reference

Genberg, Hans. "The Concept and Measurement of the World Price Level and Rate of Inflation." *J. Monetary Econ.* 3, no. 2 (1977): 231–52.

David Ricardo: A Biography. By DAVID WEATHERALL.
The Hague: Martinus Nijhoff, 1976. Pp. xi + 201. Fl 35.

This volume has little about the economics of David Ricardo but is an informative account of his life and personality. The author well summarizes the approach of the book in the preface: "Though much necessarily is said about the economist in the book, I must make plain that it is not primarily a study of his work. . . . In the arrangement of the book I have departed occasionally from a strictly chronological narrative. Instead I have attempted to present the life of David Ricardo in a number of facets, as if it were a diamond."

As the author acknowledges in the preface, he has drawn heavily upon the material in Sraffa's *The Works and Correspondence of David Ricardo.* He has also used some other sources, in particular those relating to the family background of the Ricardos, going back to Spain; the family's residence in the seventeenth century in Livorno, where they were in the coral trade; the history of the Jewish community in Amsterdam; Ricardo's religion and his association with Unitarianism; and the housekeeping details of his various homes in England.

The book reinforces the picture of Ricardo that is familiar to those who have read the correspondence and biographical material in the Ricardo *Works*: his reasonableness, his high sense of integrity, his concern for those less financially fortunate than himself—both in his own family and outside—and his efforts in behalf of religious toleration and of political reform. For the noneconomist who has read too deeply in Carlyle and Ruskin about the heartlessness of classical economists, the book should do much to correct an all too common stereotype.

Within the limits that he has set for himself, Weatherall has written an interesting account that the noneconomist as well as the economist will find good reading.

FRANK WHITSON FETTER

Hanover, New Hampshire

Index to Volume 85

Articles and Notes

Aghevli, Bijan B. Inflationary Finance and Growth 1295
Ali, Mukhtar M. Probability and Utility Estimates for Racetrack Bettors . 803
Barnett, William A. Pollak and Wachter on the Household Production Function Approach 1073
Barro, Robert J., and Friedman, James W. On Uncertain Lifetimes 843
Battalio, Raymond C.; Kagel, John H.; and Reynolds, Morgan O. Income Distributions in Two Experimental Economies 1259
Baumol, William J., and Braunstein, Yale M. Empirical Study of Scale Economies and Production Complementarity: The Case of Journal Publication 1037
Becker, Gary S.; Landes, Elisabeth M.; and Michael, Robert T. An Economic Analysis of Marital Instability 1141
Boyer, Kenneth D. Minimum Rate Regulation, Modal Split Sensitivities, and the Railroad Problem 493
Braunstein, Yale M. See *Baumol, William J.*
Bruce, Neil. The IS-LM Model of Macroeconomic Equilibrium and the Monetarist Controversy 1049
Buiter, Willem H., and Woglom, Geoffrey. On Two Specifications of Asset Equilibrium in Macroeconomic Models: A Note 395
Burmeister, Edwin. See *Turnovsky, Stephen J.*
Burns, Malcolm R. The Competitive Effects of Trust-Busting: A Portfolio Analysis 717
Calvo, Guillermo A., and Rodriguez, Carlos Alfredo. A Model of Exchange Rate Determination under Currency Substitution and Rational Expectations 617
Cardell, Nicholas Scott, and Hopkins, Mark Myron. Education, Income, and Ability: A Comment 211
Crain, W. Mark. On the Structure and Stability of Political Markets . . 829
Cropper, M. L. Health, Investment in Health, and Occupational Choice . 1273
Edwards, Franklin R. Managerial Objectives in Regulated Industries: Expense-Preference Behavior in Banking 147
Ehrenberg, Ronald G. Household Allocation of Time and Religiosity: Replication and Extension 415
Ehrlich, Isaac. Capital Punishment and Deterrence: Some Further Thoughts and Additional Evidence 741
Feldstein, Martin. The Surprising Incidence of a Tax on Pure Rent: A New Answer to an Old Question 349
Fischer, Stanley. Long-Term Contracts, Rational Expectations, and the Optimal Money Supply Rule 191

Foley, Duncan K. Reply to Buiter and Woglom 401
Freeman, Richard B. Political Power, Desegregation, and Employment of Black Schoolteachers 299
Frenkel, Jacob A., and Levich, Richard M. Transaction Costs and Interest Arbitrage: Tranquil versus Turbulent Periods 1209
Friedman, Benjamin M. Financial Flow Variables and the Short-Run Determination of Long-Term Interest Rates 661
Friedman, David. A Theory of the Size and Shape of Nations 59
Friedman, James W. See *Barro, Robert J.*
Friedman, Milton. Nobel Lecture: Inflation and Unemployment 451
Fullerton, David J. See *Rosen, Harvey S.*
Gould, J. R. Price Discrimination and Vertical Control: A Note 1063
Griliches, Zvi. Rejoinder to Cardell and Hopkins 215
Gronau, Reuben. Leisure, Home Production, and Work—the Theory of the Allocation of Time Revisited 1099
Handa, Jagdish. Risk, Probabilities, and a New Theory of Cardinal Utility 97
Hayakawa, Hiroaki, and Venieris, Yiannis. Consumer Interdependence via 1099 Reference Groups 599
Heckman, James J., and Willis, Robert J. A Beta-logistic Model for the Analysis of Sequential Labor Force Participation by Married Women 27
Higinbotham, Harlow N. See *Telser, Lester G.*
Hoffman, Emily P. The Deeper Economics of Sleeping: Important Clues toward the Discovery of Activity *X* 647
Hopkins, Mark Myron. See *Cardell, Nicholas Scott*
Jones, Ronald W., and Scheinkman, Jose. The Relevance of the Two-Sector Production Model in Trade Theory 909
Kagel, John H. See *Battalio, Raymond C.*
Keeler, Emmett B.; Morrow, Daniel T.; and Newhouse, Joseph P. The Demand for Supplementary Health Insurance, or Do Deductibles Matter? . . . 789
Keeler, Theodore E., and Small, Kenneth A. Optimal Peak-Load Pricing, Investment, and Service Levels on Urban Expressways 1
Khan, Mohsin S. The Variability of Expectations in Hyperinflations . . . 817
King, A. Thomas. Estimating Property Tax Capitalization: A Critical Comment 425
Klein, Benjamin. The Demand for Quality-adjusted Cash Balances: Price Uncertainty in the U.S. Demand for Money Function 691
Kydland, Finn E., and Prescott, Edward C. Rules Rather than Discretion: The Inconsistency of Optimal Plans 473
Landes, Elisabeth M. See *Becker, Gary S.*
Lazear, Edward. Education: Consumption or Production? 569
Levhari, David, and Mirman, Leonard J. Savings and Consumption with an Uncertain Horizon 265
Levi, Maurice D. Taxation and "Abnormal" International Capital Flows . 635
Levich, Richard M. See *Frenkel, Jacob A.*
Long, Stephen H., and Settle, Russell F. Household Allocation of Time and Church Attendance: Some Additional Evidence 409
McCabe, James L. Regional Product Price Differences and the Sectoral Distribution of Labor in Less Developed Countries 549
McCallum, Bennett T. Price-Level Stickiness and the Feasibility of Monetary Stabilization Policy with Rational Expectations 627
McCulloch, Rachel, and Yellen, Janet L. Factor Mobility, Regional Development, and the Distribution of Income 79
Machlup, Fritz. Publishing Scholarly Books and Journals: Is It Economically Viable? 217

MacRae, C. Duncan. A Political Model of the Business Cycle 239
Mehra, Y. P. Money Wages, Prices, and Causality 1227
Michael, Robert T. See *Becker, Gary S.*
Mirman, Leonard J. See *Levhari, David.*
Mohabbat, Khan A., and Simos, Evangelos O. Consumer Horizon: Further Evidence 851
Morrow, Daniel T. See *Keeler, Emmett B.*
Newhouse, Joseph P. See *Keeler, Emmett B.*
Niehans, Jürg. Exchange Rate Dynamics with Stock/Flow Interaction . . 1245
O'Driscoll, Gerald P., Jr. The Ricardian Nonequivalence Theorem 207
O'Hara, Donald J. Location of Firms within a Square Central Business District 1189
Ohlson, James A. Equilibrium in Stable Markets 859
Ohlsson, Lennart. Patterns of Engineering Trade Specialization, 1960–1970, and Sweden's Factor Abundance. 361
Phelps, Edmund S., and Taylor, John B. Stabilizing Powers of Monetary Policy under Rational Expectations 163
Pollak, Robert A., and Wachter, Michael L. Reply: "Pollak and Wachter on the Household Production Approach" 1083
Prescott, Edward C. See *Kydland, Finn E.*
Ray, Edward John. Foreign Direct Investment in Manufacturing 283
Reid, Joseph D., Jr. The Theory of Share Tenancy Revisited—Again . . 403
Reynolds, Morgan O. See *Battalio, Raymond C.*
Rodriguez, Carlos Alfredo. See *Calvo, Guillermo A.*
Rosen, Harvey S., and Fullerton, David J. A Note on Local Tax Rates, Public Benefit Levels, and Property Values 433
Rosenzweig, Mark R. The Demand for Children in Farm Households . . 123
Scheinkman, Jose. See *Jones, Ronald W.*
Settle, Russell F. See *Long, Stephen H.*
Shapley, Lloyd, and Shubik, Martin. Trade Using One Commodity as a Means of Payment 937
Shapley, L. S., and Shubik, M. An Example of a Trading Economy with Three Competitive Equilibria 873
Shiller, Robert J., and Siegel, Jeremy J. The Gibson Paradox and Historical Movements in Real Interest Rates 891
Shubik, M. See *Shapley, L. S.*
Shubik, Martin. See *Shapley, Lloyd.*
Siegel, Jeremy J. See *Shiller, Robert J.*
Simos, Evangelos O. See *Mohabbat, Khan A.*
Sjaastad, Larry A., and Wisecarver, Daniel L. The Social Cost of Public Finance 513
Small, Kenneth A. See *Keeler, Theodore E.*
Smith, Vernon L. The Principle of Unanimity and Voluntary Consent in Social Choice 1125
Stein, John Picard. The Monetary Appreciation of Paintings 1021
Stigler, George J. The Conference Handbook 441
Taylor, John B. See *Phelps, Edmund S.*
Telser, Lester G., and Higinbotham, Harlow N. Organized Futures Markets: Costs and Benefits 969
Tolley, George S., and Wilman, John D. The Foreign Dependence Question . 323
Turnovsky, Stephen J., and Burmeister, Edwin. Perfect Foresight, Expectational Consistency, and Macroeconomic Equilibrium 379
Usher, Dan. Public Property and the Effects of Migration upon Other Residents of the Migrants' Countries of Origin and Destination. . . . 1001

Venieris, Yiannis. See *Hayakawa, Hiroaki.*
Wachter, Michael L. See *Pollak, Robert A.*
Warr, Peter G. On the Shadow Pricing of Traded Commodities 865
Willis, Robert J. See *Heckman, James J.*
Wilman, John D. See *Tolley, George S.*
Wisecarver, Daniel L. See *Sjaastad, Larry A.*
Woglom, Geoffrey. See *Buiter, Willem H.*
Yellen, Janet L. See *McCulloch, Rachel.*

Reviews

Ackerman, Bruce A.; Ackerman, Susan Rose; Sawyer, James W., Jr.; and Henderson, Dale W. *The Uncertain Search for Environmental Quality.* By Larry E. Ruff 227
Barro, Robert J., and Grossman, Herschel I. *Money, Employment and Inflation.* By Benjamin M. Friedman . . . . [illegible]
Bhagwati, Jagdish N., and Srinivasan, T. N. *Foreign Trade Regimes and Economic Development: India.* By Koilpillai J. Charles 877
Calleo, David P., ed. *Money and the Coming World Order.* By Benjamin J. Cohen 651
Caves, Richard E., and Uekusa, Masu. *Industrial Organization in Japan.* By M. Bronfenbrenner 653
Cochrane, Willard W., and Ryan, Mary E. *American Farm Policy,* 1948–1973. By Don Paarlberg 879
Corden, W. M. *Trade Policy and Economic Welfare.* By Anne O. Krueger . 445
Corden, W. M., and Fels, Gerhard, eds. *Public Assistance to Industry: Protection and Subsidies in Britain and Germany.* By Edward Tower . . . 1092
Dam, Kenneth W. *Oil Resources: Who Gets What How?* By Colin Robinson 882
Dixit, A. K. *The Theory of Equilibrium Growth.* By A. Asimakopulos 1311
Evans, Alan W. *The Economics of Residential Location.* By Richard E. Schuler 655
Freeman, Richard B. *The Overeducated American.* By Cotton M. Lindsay . 1094
Furubotn, Eirik G., and Pejovich, Svetozar, eds. *The Economics of Property Rights.* By Albert Breton 229
Goodhart, C. A. E. *Money, Information and Uncertainty.* By Andrew J. Policano 231
Hymer, Stephen Herbert. *The International Operations of National Firms: A Study of Direct Foreign Investment.* By Thomas G. Parry 1096
Kain, John F., and Quigley, John M. *Housing Markets and Racial Discrimination.* By Hugh O. Nourse 885
Meade, J. E. *The Just Economy.* By Neil Bruce 1312
Mikdashi, Zuhayr. *The International Politics of Natural Resources.* By Arturo C. Porzecanski 658
Rutledge, John. *A Monetarist Model of Inflationary Expectations.* By Robert J. Shiller 448
Schneider, Erich. *Joseph A. Schumpeter: Leben und Werk eines Grossen Sozialökonomen (Life and Work of a Great Social Scientist).* Translated and introduced by W. E. Kuhn. By Gottfried Haberler 660
Smith, Adam. *The Wealth of Nations.* Edited by R. H. Campbell and A. S. Skinner; W. B. Todd, textual editor. By George J. Stigler 235

Stem, Carl H.; Makin, John M.; and Logue, Dennis E. *Eurocurrencies and the International Monetary System*. By HANS GENBERG 1314
Tendler, Judith. *Inside Foreign Aid*. By G. K. HELLEINER 236
Tyler, William G. *Manufactured Export Expansion and Industrialization in Brazil*. By JOSÉ L. CARVALHO 888
Weatherall, David. *David Ricardo: A Biography*. By FRANK WHITSON FETTER 1317

The Department of Education of the University of Chicago announces that it is seeking candidates for **two positions** in **Comparative Education** for **Fall 1978**. The candidates should have had their training and experience in Anthropology, Economics, Political Science, or Sociology and have a major commitment to such areas of study as educational change, social and economic development, educational finance, comparative organizations, or social stratification. Also desirable are specialization in Latin American or African studies, and interests in policy analysis and planning.

Appointments are open as to rank and salary. The University of Chicago is an equal opportunity employer; members of minority groups and women are encouraged to apply. Applicants should include a curriculum vitae with their application.

Write to: **Professor Robert Dreeben**
Department of Education
University of Chicago
5835 South Kimbark Avenue
Chicago, Illinois 60637

Cities and Frontiers in Brazil
Regional Dimensions of Economic Development
Martin T. Katzman

Brazil bases its aspirations for world-power status on its ability to dominate and reshape its vast space. Mr. Katzman focuses here on Brazil's diverse policies for controlling frontier settlement, regional integration, and urbanization — the creation of growth poles, development highways, and fiscal incentives — in a thorough analysis of the neglected spatial dimensions of economic development. This book is unique both in its interdisciplinary approach, drawing upon materials from anthropology, political science, history, geography, and economics, and in its full analysis of the interrelationships between urban and rural development. $16.50

HARVARD UNIVERSITY PRESS
Cambridge, MA 02138